Academic Capability Mapping

King Tao Ng

Supervisor: Roman Marchant Matus
Chao Sun
Philip Henville

Faculty of Engineering and Information Technologies
University of Sydney

A thesis submitted in partial fulfilment of requirements for the degree of
Master of Data Science

November 2018

I would like to dedicate this thesis to my wife, Joanna.

The Section 3.1.2 and 3.1.3 of this thesis is published as [31]. I co-designed the study with Chao Sun, Roman Marchant Matus, and Philip Henville.

Declaration

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes. I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

King Tao Ng
November 2018

Acknowledgements

This project is funded by the Faculty of Engineering and Information Technologies, and Philip Henville appoints me to turn the proposed solution into a product. Roman Marchant Matus and Chao Sun recommend a hierarchical structure of Word Mover's Distance. Chao Sun proposes the Word Mover's Distance Variant using Latent Dirichlet Allocation, and Joel Nothman introduces the Cosine Similarity as a benchmark.

Abstract

We analyse a sample of journals written by academics from the Faculty of Engineering and Information Technologies in order to find potential matches for research collaborations. Using the Hierarchical Word Mover's Distance, we compute distances among any pairs of academics based on their journal titles and rank distances accordingly. Shorter distances are; potentially closer pairs of researchers' profile pages are. Using calculated distances as edges and researchers as nodes, we generate network visualisations and evaluate them against known co-authorships. Because of the Word Mover's Distance's runtime performance, the Latent Dirichlet Allocation is also examined for grouping titles into topic clusters prior to computing distances. The proposed model outperforms its variant and a benchmark with regard to identifying existing co-authorships in a small dataset. Nevertheless, as dataset grows, all models become comparable. Using a calculated distance metric as a predictor variable along with others extracted from researchers' profile pages, we use generalised linear models to find out distances among pairs of researchers and departments affect co-authorships the most. Provided a pair from the same department, we use a generalised linear mixed models to conclude Australian Centre for Field Robotics or Electrical & Information Engineering is more significant than other departments.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Objectives

To estimate academic capability, and explore potential collaboration through researcher network analysis and visualisation is a key insight for various areas of an organisation. This thesis intends to develop network visualisations among researchers using various distance metrics. These visualisations can be used as an exploratory tool to encourage a researcher expanding her network not only to people within the same fields, but also to cross-disciplinary researchers with different skill sets. Furthermore, using a calculated distance metric as a predictor variable along with others extracted from researchers' profile pages, we use generalised linear models to determine whether any predictor variables affect co-authorships and estimate their magnitudes, if any. In other words, our aims are to

- Develop and evaluate proposed methods for calculating distance metrics among researchers,
- Design and implement visuals using calculated distance metrics, and
- Determine any predicator variables affect co-authorship counts using generalised linear models.

Ultimately, a tool we develop will be used in the Faculty of Engineering and Information Technologies for researcher recommendations. That is, the tool recommends whom you may want to collaborate with based on publications and researchers' profile pages.

1.2 Research Questions

Quite often people in the Faculty of Engineering and Information Technologies want to find out what expertises researchers have so that they can approach them for collaborations. Currently the way to achieve this is through profile pages. Therefore, they have requested for a researcher recommendations system to be built. They have also asked the following questions:

- Can we match researchers for potential future collaborations based on their publications?
- While publications are a critical factor for researcher recommendations, arguably, social factors are equally important. Is it possible to incorporate social factors to improve researcher recommendations?

1.3 Literature Review

Many institutes have implemented such collaborative networks on the researchers' profile pages based on co-authorships or citations [28]. However, such visualisation does not show much value to the researchers because they already know whom they have collaborated with in the past. Apart from co-authorships, Gollapalli et al. [17] computed similarity by including profiles extracted from their publications and academic homepages. On the contrary, Xu et al. [34] concluded researcher collaboration systems had been thus far broadly classified into two categories for the past few years – Semantic Analysis and Social Relations. Hence, they proposed a two-layer network approach that integrated the two together. In our view, this network becomes more complex when considering more concepts and researchers. While our aim is quite similar to Xu et al. [34], the approach is completely different. We integrate the two by means of the generalised linear models.

1.4 Outline

The thesis is organized into 7 chapters. This Chapter details the objectives as well as the literature review. The review continues to the Chapter 2 where it examines a few statistical techniques to compute a distance between two documents. We also explain their shortcomings. Eventually, it leads us to the Word's Mover Distance and Latent Dirichlet Allocation. We present proposed methods and a benchmark in the Chapter 3.

We then explain the experimental set-up and provide results in the Chapter 4. In the Chapter 5, we walk through a couple of visuals as well as their implementations. We walk through the generalised linear models and generalised linear mixed models in the Chapter 6. In the last Chapter, we provide suggestions to implementation, and discuss future works.

Furthermore, an sample of data is shown in the Appendix A. The Appendix B lists down hardware and software required in order to run the application. The Appendix C shows milestones.

Chapter 2

Mathematical Background

2.1 Introduction

Representing a distance between two researchers' profile pages can be simplified as a distance between two journals written individually by them. Once distances among all pairs of their journals are computed, a distance between two researchers' profile pages can be somewhat calculated.

One of common ways to represent a journal is the *Term Frequency* (TF) [24]. The TF can be graphically plotted as a histogram where the x-axis represents distinct words and y-axis represents normalised frequencies in which the total area is summed to 1. For instance, let us take the abstract *Fundamental value: a category in transformation*[1] from Prof. Richard Bryan as an example:

'Fundamental value' is a canonical category in both Marxian and neo-classical economics. In application to finance and financial crisis, it is laden with complexity. For Marxists, it has underscored a focus on the distinctions between production and circulation, real and fictitious capital. These debates have dominated Marxian responses to the financial crisis. In mainstream finance and economics, the term has undergone transformation and historically driven adaptation. Marxian analysis could fruitfully follow this lead. This paper identifies that transformation as an expression of capital's transforming calculation project. As capital becomes more liquid, the concept of fundamental value itself must embrace liquidity, yet, in embracing liquidity, fundamental value loses its established definitive capacity.

After removing punctuations and converting words into lower cases, the corresponding histogram is shown in the Fig. 2.1. Similarly, the abstract *Current Emotion*

[1] http://www.tandfonline.com/doi/abs/10.1080/03085147.2012.718625

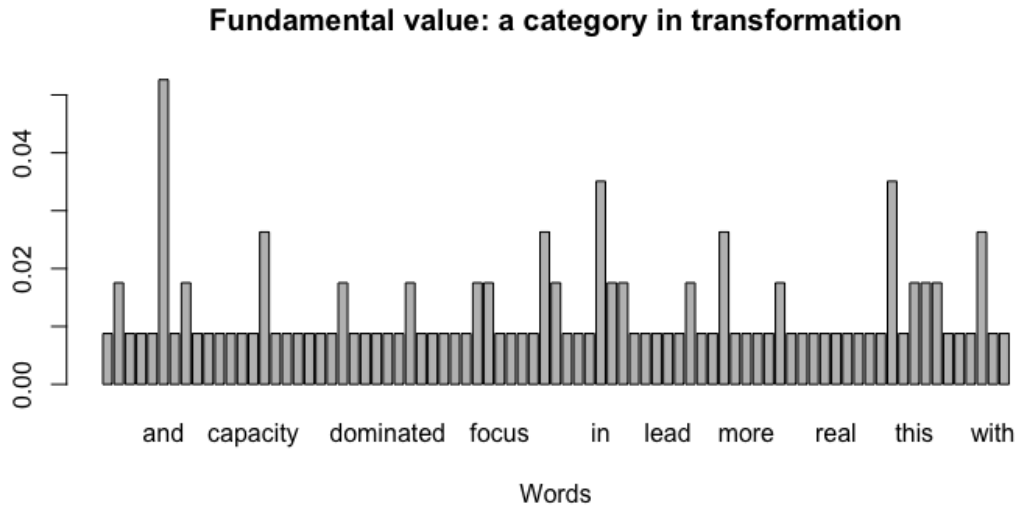

Fig. 2.1 Word frequencies of *Fundamental value: a category in transformation*

Research in Philosophy[2] from Prof. Paul E. Griffiths can be plotted as a histogram in the Fig. 2.2:

There remains a division between the work of philosophers who draw on the sciences of the mind to understand emotion and those who see the philosophy of emotion as more self-sufficient. This article examines this methodological division before reviewing some of the debates that have figured in the philosophical literature of the last decade: whether emotion is a single kind of thing, whether there are discrete categories of emotion, and whether emotion is a form of perception. These questions have been addressed by both sides of the methodological divide and the integration of these two approaches would have clear benefits.

2.2 χ^2 Test

To measure similarity between two histograms, the simplest way is the χ^2 Test [30, 29]. Suppose we know m_i is the occurrence of the ith word from the journal m and the n_i is the frequency of the ith word from the journal n. Mathematically, the χ^2 Test can be denoted as

$$\chi^2 = \sum_{i=1}^{N} \frac{(m_i - n_i)^2}{n_i}, \tag{2.1}$$

[2]http://journals.sagepub.com/doi/abs/10.1177/1754073912468299

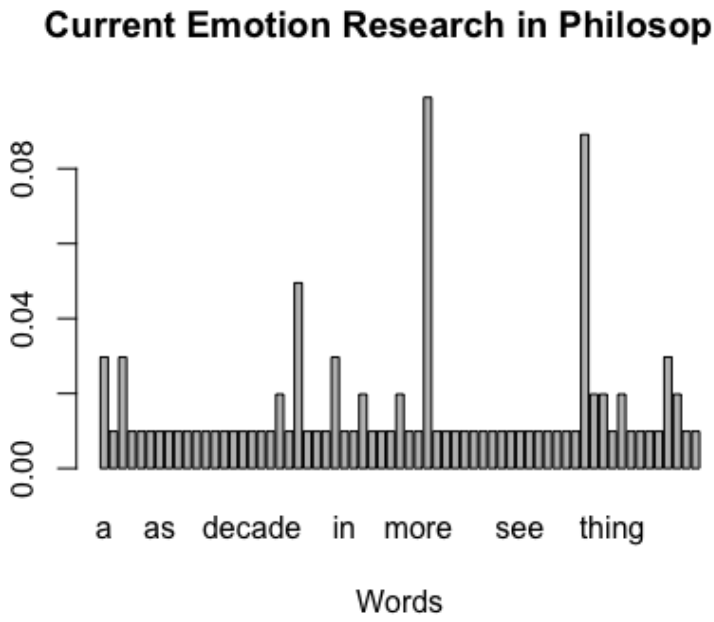

Fig. 2.2 Word frequencies of *Current Emotion Research in Philosophy*

where N is the total number of words from both m and n. A large value of χ^2 indicates the null hypothesis where two journals come from the same distribution is less likely. However, it fails to consider a situation where two journals do not share the same syntactic form, but conveys nearly the same information [23], such as *Obama speaks to the media in Illinois* and *The President greets the press in Chicago*. The χ^2 Test would almost certainly reject the null hypothesis in this situation.

2.3 Kullback-Leibler Divergence

Alternatively, if we approximate from one histogram to another, is it possible to measure how much information is expected to lose during the approximation? Kullback-Leibler Divergence simply measures such metrics [30, 22, 27]. Again, we have two journals, namely m and n respectively. m_i is the occurrence of the ith word from the journal m and the n_i is the frequency of the ith word from the journal n.

$$D(m \parallel n) = \sum_{i=1}^{N} m_i(\log(m_i) - \log(n_i)) \tag{2.2}$$

Unfortunately, this approach suffers the same problem as the χ^2 Test. Also, the divergence is not a distance because divergence is not symmetric. In other words,

$$D(m \parallel n) \neq D(n \parallel m)$$

2.4 Word Mover's Distance

Because both χ^2 Test and Kullback-Leibler Divergence adopt a bin-by-bin approach, only the same words can be considered [30]. How do we compute similarity between two words, particularly words do not share the same syntactic form, but convey nearly the same semantic meaning?

2.4.1 Word Embedding

Matt et al. [23] proposed to solve this problem using the word embedding. The *word2vec* model is one way to produce the word embedding. The model represents words in the vector space in which words that are semantically similar in the corpus are located close to each other. More importantly, distances are symmetric. For instance, we can see words that are the most semantically similar with the word *Obama*, and visualise the embeddings using t-SNE [32, 5] in the Fig. 2.3.

A distance between two words can be calculated using the Euclidean distance. For instance, *Obama* and *Barack* are located approximately at $\vec{v} = (0.000175, 0.00004)$ and $\vec{u} = (0.00015, -0.000025)$ respectively, so the distance between them, or the length between a vector $\vec{v}$ and a vector $\vec{u}$, can be computed as follows:

$$\|\vec{v} - \vec{u}\| = \sqrt{(0.000175 - 0.00015)^2 + (0.00004 + 0.000025)^2} = 6.9641e^{-05}$$

2.4.2 Earth Mover's Distance

However, a journal more likely has more than one word, so it is critical to take word counts into consideration. To circumvent this problem, Matt et al. [23] recommended using the Earth Mover's Distance. Let us assume we have two documents called m and n respectively whereas m has the following words: *cat*, *cat*, *dog*, *human* and *tiger*; n has *kitten*, *cat*, *kitten*, *man* and *lion*. Corresponding histograms are plotted in the Fig. 2.4. Again, the area in the histogram is summed to 1. To work out a distance between these two documents, each bin is chopped into smaller chunks, which are then

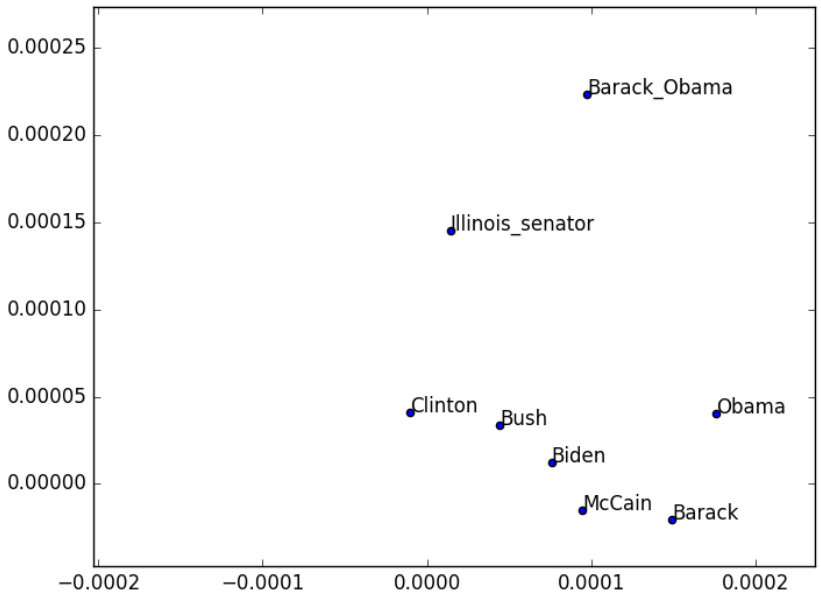

Fig. 2.3 A word2vec of the word *Obama*

moved from one histogram to another in order to fill up the corresponding area (See the Fig. 2.5).

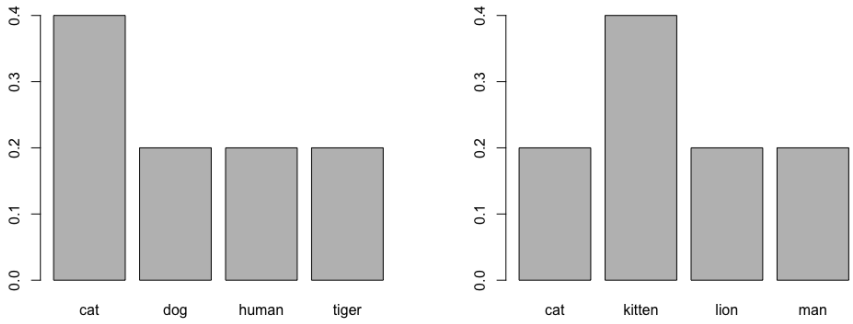

Fig. 2.4 Histograms of word counts for m and n respectively

Since we know distances among any pair of words from the word2vec matrix in the Table 2.1, a chunk is moved such that its moving distance is minimised. We have adopted the notation from Matt et al. [23] the word2vec matrix is called $c(i, j)$, which is a distance between word i to word j. For instance, the cat bin in m is divided up into two chunks where the bottom one is moved to the cat bin in n because c(cat,

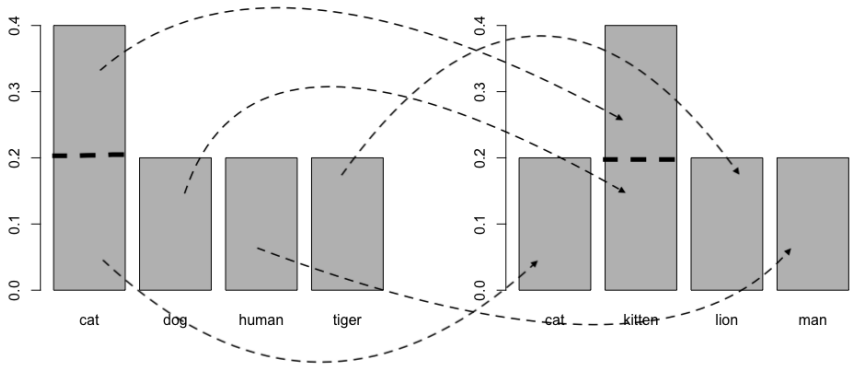

Fig. 2.5 Histograms of word counts for m and n respectively with arrows

cat) = 0, and the top one is moved to the kitten bin in n because c(cat, kitten) = 0.1 in which its distance is the second shortest from cat. The size of a single chuck is 0.25 because all chucks must be summed to 1. A distance for the first bin in m is therefore calculated by adding all distances up, in other words, $0.25 \times c$(cat, cat) $+ 0.25 \times c$(cat, kitten) = 0.025. Other bins follow the same principle, and a distance between these two documents is simply to add up all bin distances. Together with the word embedding, this approach is called the Word Mover's Distance [23].

The Word Mover's Distance can be mathematically formulated as follows:

$$\begin{aligned} \underset{t_{i,j} \geq 0}{\text{Minimize}} \quad & \sum_{i=1}^{M} \sum_{j=1}^{N} t_{i,j} c(i,j) \\ \text{Subject to} \quad & \sum_{i=1}^{M} t_{i,*} = 1. \\ & \sum_{i=1}^{N} t_{*,j} = 1. \end{aligned} \tag{2.3}$$

where $t_{i,j}$ is the size of a single chunk that moves from word i to word j. However, when a chunk is not allowed to move, $t_{i,j} = 0$, in other words, a chunk in the corresponding histogram has been filled. The aim is to minimize the total distance cost. The constraint conditions say all chunks must be added up to 1 in both m and n respectively.

	cat	**dog**	**human**	**tiger**
cat	0	0.5	1.2	0.8
kitten	0.1	0.6	1.5	0.9
lion	0.9	0.8	1.6	0.5
man	1.3	1.6	0.2	1.4

Table 2.1 Distances from word2vec matrix

2.5 Cosine Similarity

The Word Mover's Distance is one way to compute a distance between two journals, but it is not the only way. In fact, the Cosine Similarity [25] can also quantify similarity between two journals m and n:

$$sim(m, n) = \frac{\vec{m}\vec{n}}{\|\vec{m}\|\|\vec{n}\|}, \tag{2.4}$$

where $\vec{m}$ and $\vec{n}$ are vector representations of m and n respectively. Apart from the TF introduced in the Section 2.1, the *Term Frequency–Inverse Document Frequency* (TF–IDF) [24] is another way, or arguably more accurate way, to represent the journal. $\vec{m}$ and $\vec{n}$ have to be a vector of either one of these two representations. Compared with the word embedding, these representations lose the relative ordering of words. For instance, *Mary is more beautiful than Jane* and *Jane is more beautiful than Mary.* They have identical representations, but clearly they do not have the same semantic meaning.

2.6 Latent Dirichlet Allocation

One critical shortcoming of the Word Mover's Distance is the runtime performance. As a number of words in a journal grows bigger, a pairwise comparison among two journals definitely deteriorates the runtime performance further. To reduce a number of comparisons, we group journals into topic clusters prior to computing distances. The Latent Dirichlet Allocation [3, 6] describes a set of words collected from journals in term of a mixture of a small number of topics.

Chapter 3

Methodology

3.1 Proposed Models

Having explained the mathematical background, we now present the Hierarchical Word Mover's Distance[1] for learning a distance between two researchers, and its variant using the Latent Dirichlet Allocation for grouping journals into topic clusters prior to computing distances. Finally, we also compare proposed models with the baseline model – Cosine Similarity in both TF and TF–IDF representations[2].

3.1.1 Hierarchical Word Mover's Distance

In the preprocessing stage, English stopwords must be removed from journals prior to any computation. In the first stage, we can calculate distances among any pairs of journals as demonstrated in the Section 2.4. In the second stage, a distance between two researchers' profile pages is calculated using the same idea. For instance, let us assume we have two individuals, namely John and Paul, who publish journals A1, A2 and A3, and A1, A4 and A5 respectively as shown in the Fig. 3.1. As you may have noticed, a bin size is always the same simply because each journal is weighted equally. The area is summed to 1.

To find out a distance between John and Paul, we again move each bin from one histogram to another such that its moving distance is minimised. Because the journal A1 is co-authored, a distance to move from A1 to A1 itself is obviously zero. If a distance between A2 and A4 is shorter than between A2 and A5, a distance between

[1]Roman Marchant Matus and Chao Sun recommend a hierarchical structure of Word Mover's Distance.

[2]Chao Sun proposes the Word Mover's Distance Variant using Latent Dirichlet Allocation, and Joel Nothman introduces the Cosine Similarity as a benchmark.

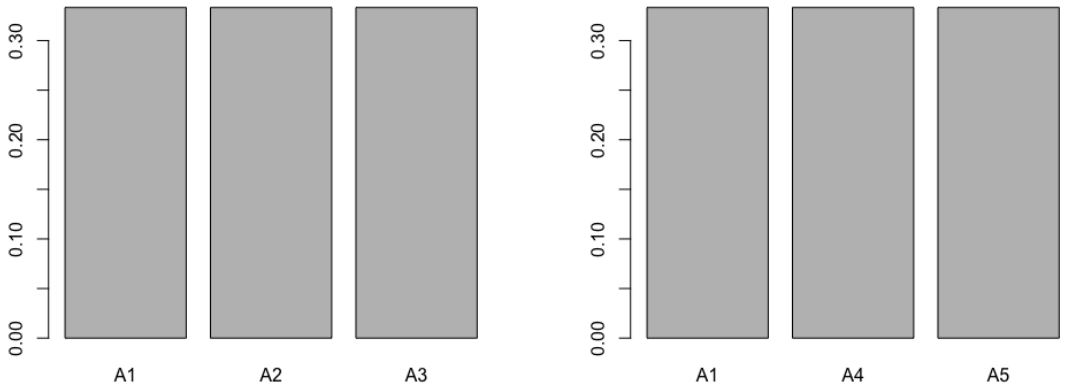

Fig. 3.1 Histograms of journals for John and Paul respectively

John and Paul is therefore a distance between A2 and A4 plus a distance between A3 and A5.

The model runs for all pairs. Calculated distances are sorted in an ascending order. A pair with the shortest distance is ranked on the top; a pair with the furthest distance is ranked on the bottom.

3.1.2 Hierarchical Word Mover's Distance Variant

Similar to the model introduced in the Section 3.1.1, this model maps out journals into topic distributions based on the Latent Dirichlet Allocation before computing distances using the above method. Using the perplexities, Chao et al. [31] suggested setting a number of topics to 41 would yield an optimal performance. However, 40 has been arbitrarily chosen in the experiment.

3.1.3 Cosine Similarity

The baseline model, or Cosine Similarity, is introduced for benchmarking. The model is created by concatenating all journals into a big document for each researcher, and then we compare two documents from two different researchers each time. The documents must be represented in the either TF or TF–IDF representation. When two documents are identical, similarity yields a score of 1; When two documents are totally different, similarity yields a score of 0. Similarities are therefore sorted in a descending order. That is, a pair on the top is the most similar.

3.2 Data

Data for the study is extracted from Google Scholarly [13] and Scopus [20] based on journals from 18 researchers who work for the Faculty of Engineering and Information Technologies. Furthermore, a total number of journals across all researchers is 2,622. The Table 3.1 shows a breakdown for each researcher.

Researcher	**Number of journals published**
Xiangyuan Carl Cui	70
Xiaozhou Liao	173
Qing Li	247
Salah Sukkarieh	160
Simon Ringer	347
Julie Cairney	165
Ian Manchester	84
Zongwen Liu	124
Yuan Chen	112
Jun Huang	100
Luming Shen	112
Itai Einav	141
Daniel Dias Da Costa	89
Branka Vucetic	285
Judy Kay	107
Jinman Kim	98
Dacheng Tao	104
Dong Xu	104
Total	**2,622**

Table 3.1 Number of journals published for each researcher

Data, which comes as the `JSON` format, includes titles, abstracts, authors, publishers, years and etc. An sample of data is shown in the Appendix A. The Table 3.2, which has been sorted in a descending order on the Co-authorships, shows a breakdown of co-authorships that occurs for each pair of researchers. For example, Simon Ringer and Julie Cairney co-write papers 59 times. This table is also used as the ground truth when evaluating proposed models.

Researcher	**Researcher**	**Co-authorships**
Simon Ringer	Julie Cairney	59
Xiaozhou Liao	Simon Ringer	50
Simon Ringer	Zongwen Liu	41
Xiangyuan Carl Cui	Simon Ringer	37
Dacheng Tao	Dong Xu	8
Luming Shen	Itai Einav	7
Zongwen Liu	Jun Huang	5
Xiaozhou Liao	Julie Cairney	5
Xiangyuan Carl Cui	Zongwen Liu	5
Luming Shen	Daniel Dias Da Costa	2
Xiaozhou Liao	Jun Huang	2
Xiaozhou Liao	Yuan Chen	2
Zongwen Liu	Luming Shen	1
Simon Ringer	Luming Shen	1
Xiaozhou Liao	Zongwen Liu	1
Zongwen Liu	Yuan Chen	1
Xiangyuan Carl Cui	Julie Cairney	1
Xiangyuan Carl Cui	Xiaozhou Liao	1
Julie Cairney	Daniel Dias Da Costa	1
Salah Sukkarieh	Jinman Kim	1
Daniel Dias Da Costa	Branka Vucetic	1

Table 3.2 Co-authorship occurrences among pairs of researchers

Chapter 4

Experiments

4.1 Experimental Set-up

An experiment is set up in the way journal titles, rather than journal themselves, are fed into the models, including the baselines (See the Section 7.1 for an explanation). Titles are accompanied by researcher's names, so the models know exactly who–writes–what–journals. The experiment is run multiple times with gradually increasing sample sizes, starting from 90 journals all the way to 2,622. A journal where researchers co-write is removed from the dataset. The idea is to compute a distance or similarity matrix such that a pair with the shortest distance or highest similarity is ranked on the top and a pair with the furthest distance or lowest similarity is ranked on the bottom.

We have three possible hypotheses, which are the Hierarchical Word Mover's Distance, Hierarchical Word Mover's Distance Variant, and Cosine Similarity (both TF and TF–IDF). The goal is to find the best hypothesis that approximates the co-authorship counts. In other words, a closer distance the pair is; more co-authorships it should be. This is the unsupervised learning in which the ground truth is not provided when training the models.

4.2 Ranked Evaluations

Distances or similarity scores that the models produce are artificial values. That is, one cannot prove or disprove them. For instance, if a model showed the distance between Simon Ringer and Julie Cairney was 0.856, it would be difficult to convince they were precisely this close. In statistics, this is called a *Latent Variable*. Fortunately, distances and scores do preserve the ordering, so it is possible to evaluate their ranks [25, 14].

4.2.1 Precision, Recall and F1-Score

The three most commonly used evaluation measures are precision, recall and F1-score [14]. Thankfully, we can make use of them with a slight modification to suit our needs. They are now calculated for each "rank" (i.e. the top 1, top 2, and etc. results). Hence, for the top n results, they are defined as follows:

$$Precision_n = \frac{\text{a number of retrieved pairs that have at least one co-authorship}}{\text{a number of retrieved pairs}} \tag{4.1}$$

$$Recall_n = \frac{\text{a number of retrieved pairs that have at least one co-authorship}}{\text{a number of pairs that have at least one co-authorship}} \tag{4.2}$$

$$F1_n = 2 \times \frac{Precision_n \times Recall_n}{Precision_n + Recall_n} \tag{4.3}$$

Perhaps the best way to illustrate these measures is an example. The Table 4.1, which has been sorted based on the Distance in an ascending order, shows the top 30 results from the Hierarchical Word Mover's Distance on the sample size of 681 along with precision, recall and F1-score. At the rank 3, in the case of precision, the denominator is 3 because of 3 retrieved pairs; the numerator is also 3 as all retrieved pairs thus far have at least one co-authorship, so $Precision_3 = \frac{3}{3} = 1$. For recall, the numerator is still 3 whereas the denominator is 6 because we have 6 pairs of researchers in total that have at least one co-authorship. Hence, $Recall_3 = \frac{3}{6} = 0.5$. Similarly, at the rank 5, $Precision_5 = \frac{4}{5} = 0.8$ whereas $Recall_5 = \frac{4}{6} = 0.667$. For a quick reference, we have also calculated precisions, recalls and F1-scores for all the ranks.

Rank	Researcher	Researcher	Distance	Co-authorships	Precision	Recall	F1
1	Xiangyuan Carl Cui	Simon Ringer	0.900355001	5	1	0.166666667	0.285714286
2	Simon Ringer	Zongwen Liu	0.933017757	4	1	0.333333333	0.5
3	Dacheng Tao	Dong Xu	0.933154214	1	1	0.5	0.666666667
4	Xiaozhou Liao	Simon Ringer	0.963705754	2	1	0.666666667	0.8
5	Simon Ringer	Julie Cairney	0.985564883	0	0.8	0.666666667	0.727272727
6	Xiangyuan Carl Cui	Zongwen Liu	0.996550449	2	0.833333333	0.833333333	0.833333333
7	Luming Shen	Daniel Dias Da Costa	1.004631724	1	0.857142857	1	0.923076923
8	Xiaozhou Liao	Luming Shen	1.007636645	0	0.75	1	0.857142857
9	Luming Shen	Itai Einav	1.022966406	0	0.666666667	1	0.8
10	Jinman Kim	Dong Xu	1.031145657	0	0.6	1	0.75
11	Zongwen Liu	Yuan Chen	1.031183709	0	0.545454545	1	0.705882353
12	Jinman Kim	Dacheng Tao	1.032885074	0	0.5	1	0.666666667
13	Zongwen Liu	Jun Huang	1.038739283	0	0.461538462	1	0.631578947
14	Yuan Chen	Jun Huang	1.044062809	0	0.428571429	1	0.6
15	Simon Ringer	Luming Shen	1.05101759	0	0.4	1	0.571428571
16	Qing Li	Luming Shen	1.053672652	0	0.375	1	0.545454545
17	Xiaozhou Liao	Zongwen Liu	1.057346406	0	0.352941176	1	0.52173913
18	Xiaozhou Liao	Julie Cairney	1.05941895	0	0.333333333	1	0.5
19	Simon Ringer	Yuan Chen	1.059533723	0	0.315789474	1	0.48
20	Xiaozhou Liao	Yuan Chen	1.06021927	0	0.3	1	0.461538462
21	Zongwen Liu	Luming Shen	1.064137254	0	0.285714286	1	0.444444444
22	Xiangyuan Carl Cui	Julie Cairney	1.064316869	0	0.272727273	1	0.428571429
23	Julie Cairney	Luming Shen	1.066035073	0	0.260869565	1	0.413793103
24	Julie Cairney	Zongwen Liu	1.066087232	0	0.25	1	0.4
25	Yuan Chen	Luming Shen	1.066920931	0	0.24	1	0.387096774
26	Itai Einav	Daniel Dias Da Costa	1.068429808	0	0.230769231	1	0.375
27	Qing Li	Daniel Dias Da Costa	1.068629658	0	0.222222222	1	0.363636364
28	Simon Ringer	Jun Huang	1.072038616	0	0.214285714	1	0.352941176
29	Xiangyuan Carl Cui	Jun Huang	1.079721156	0	0.206896552	1	0.342857143
30	Qing Li	Itai Einav	1.081454855	0	0.2	1	0.333333333

Table 4.1 Precision, recall and F1-score of the Hierarchical Word Mover's Distance for 681 journals

4.2.2 Precision–Recall Curve

Unsurprisingly these measures can be plotted together, as shown in the Fig. 4.1, for all the ranks. Noticeably, precision is non-increasing whereas recall is non-decreasing. On the other hand, the F1-score is a harmonic mean in which its value is closer to the lower of the two. Furthermore, the F1-score is maximised at the rank 7 when both precision and recall are high.

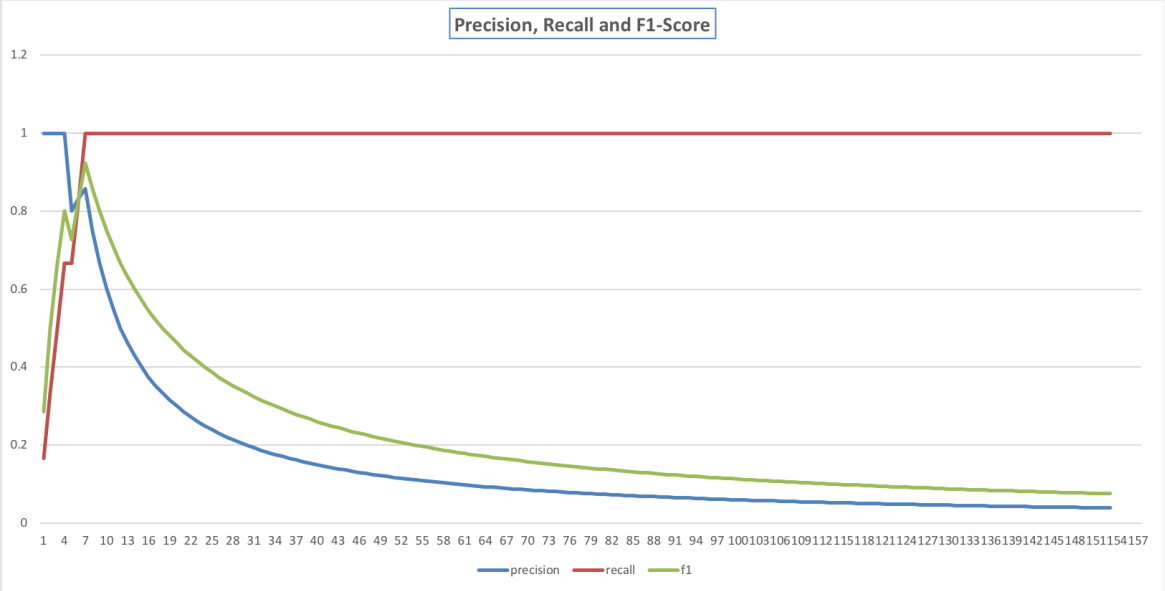

Fig. 4.1 Precision, recall and F1-score of the Hierarchical Word Mover's Distance for 681 journals

While the figure is interesting, it does not indicate how well it performs compared with other hypotheses. To address this issue, precision and recall must be plotted against each other to produce the *Precision-Recall Curve* [14] as shown in the Fig. 4.2 in which the x-axis is recall and the y-axis is precision.

Equally the same curve can be plotted for Cosine Similarity with the TF–IDF representation on the same dataset. If we overlay the two Precision-Recall Curves, we will have the Fig. 4.3. The red curve is the Hierarchical Word Mover's Distance; the blue curve is Cosine Similarity. The red curve almost always sits on the top of the blue one except toward the end when both share the same values. It is suggested the Hierarchical Word Mover's Distance is never worse than the Cosine Similarity. In fact, for the top 13 results, the Hierarchical Word Mover's Distance outperforms compared with the baseline.

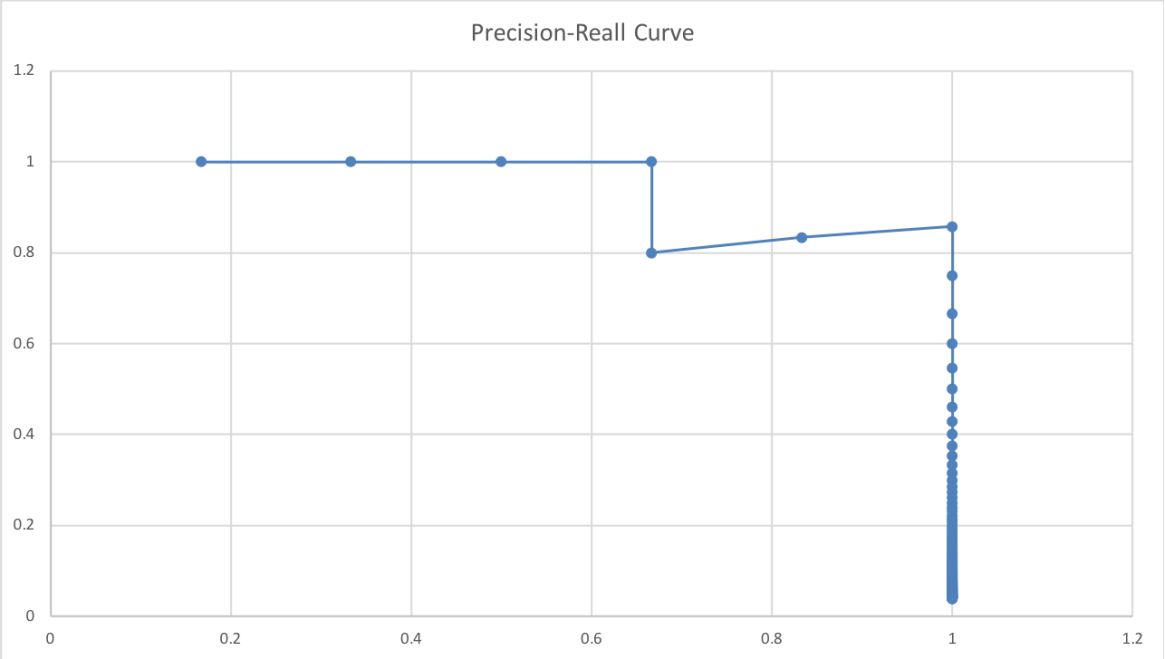

Fig. 4.2 Precision-Recall Curve of the Hierarchical Word Mover's Distance for 681 journals

4.2.3 Mean Average Precision

The Precision–Recall Curve is acceptable, but sometimes all we want is a single number that summarises the performance of hypothesis. The *Average Precision* (AP) [14] is a good candidate. On the other hand, taking an average of recall is a bad idea as we can always increase recall by returning more pairs of researchers.

To calculate the AP, we add precisions up whenever there is at least one co-authorship, and divide by the total number of pairs that have at least one co-authorship. Using the Table 4.1 as an example, for the Hierarchical Word Mover's Distance, the AP can be calculated as follows:

$$AP = \frac{\sum_{i=1}^{2,3,4,6,7} Precision_i}{6} = 0.948$$

For the sake of completeness, the AP for Cosine Similarity is 0.439, which is consistent to the Fig. 4.3, on the same dataset. The *Mean Average Precision* (MAP) [25, 14] is simply an average of this measure across all performed experiments.

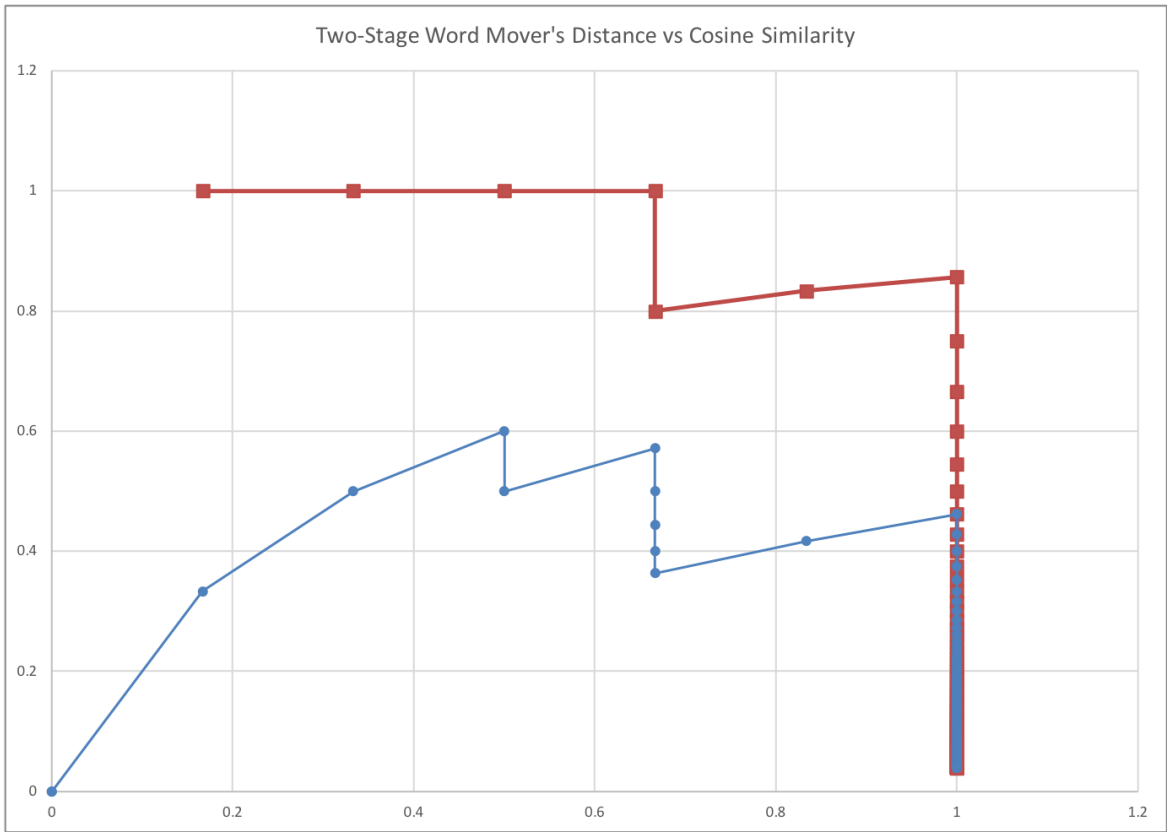

Fig. 4.3 Two Precision-Recall Curves for both the Hierarchical Word Mover's Distance and Cosine Similarity

4.2.4 Normalised Discounted Cumulative Gain

We have thus far considered if a pair has any co-authorships. In other words, a pair that has 1,000 co-authorships is treated as the same as a pair that has only 1 co-authorship. Clearly, a pair with more co-authorships should be ranked higher than a pair with less co-authorships. The *Discounted Cumulative Gain* (DCG) [25, 1] is designed to take co-authorships into account when evaluating a rank:

$$DCG = \sum_{i=1}^{N} \frac{rel_i}{\log_2(i+1)}, \tag{4.4}$$

where i is a current rank, N is a total rank, and rel is a number of co-authorships at rank i. However, this value is usually normalised to yield a value between 0 and 1, so the *Normalised Discounted Cumulative Gain* (NDCG) [25, 1] is more desirable. To do so, we have to compute the *Ideal Discounted Cumulative Gain* (IDCG) [25, 1]. The IDCG produces the maximum possible DCG by sorting co-authorships in a descending order.

$$IDCG = \sum_{i=1}^{N} \frac{rel_i}{\log_2(i+1)}, \tag{4.5}$$

where $rel_i \geq rel_{i+1}$. Hence, the NDCG is defined as follows:

$$NDCG = \frac{DCG}{IDCG} \tag{4.6}$$

The Table 4.2, which has been sorted based on the Distance in an ascending order, shows the top 30 results from the Hierarchical Word Mover's Distance on the sample size of 681 along with DCG and IDCG. From the Equation 4.6, the NDCG is simply a sum of individual DCG cells divided by a sum of individual IDCG cells, which is 0.981. Nonetheless, the NDCG in case of the Cosine Similarity is 0.571.

Rank	Researcher	Researcher	Distance	Co-authorships	DCG	IDCG
1	Xiangyuan Carl Cui	Simon Ringer	0.900355001	5	5	5
2	Simon Ringer	Zongwen Liu	0.933017757	4	2.523719014	2.523719014
3	Dacheng Tao	Dong Xu	0.933154214	1	0.5	1
4	Xiaozhou Liao	Simon Ringer	0.963705754	2	0.861353116	0.861353116
5	Simon Ringer	Julie Cairney	0.985564883	0	0	0.386852807
6	Xiangyuan Carl Cui	Zongwen Liu	0.996550449	2	0.712414374	0.356207187
7	Luming Shen	Daniel Dias Da Costa	1.004631724	1	0.333333333	0
8	Xiaozhou Liao	Luming Shen	1.007636645	0	0	0
9	Luming Shen	Itai Einav	1.022966406	0	0	0
10	Jinman Kim	Dong Xu	1.031145657	0	0	0
11	Zongwen Liu	Yuan Chen	1.031183709	0	0	0
12	Jinman Kim	Dacheng Tao	1.032885074	0	0	0
13	Zongwen Liu	Jun Huang	1.038739283	0	0	0
14	Yuan Chen	Jun Huang	1.044062809	0	0	0
15	Simon Ringer	Luming Shen	1.05101759	0	0	0
16	Qing Li	Luming Shen	1.053672652	0	0	0
17	Xiaozhou Liao	Zongwen Liu	1.057346406	0	0	0
18	Xiaozhou Liao	Julie Cairney	1.05941895	0	0	0
19	Simon Ringer	Yuan Chen	1.059533723	0	0	0
20	Xiaozhou Liao	Yuan Chen	1.06021927	0	0	0
21	Zongwen Liu	Luming Shen	1.064137254	0	0	0
22	Xiangyuan Carl Cui	Julie Cairney	1.064316869	0	0	0
23	Julie Cairney	Luming Shen	1.066035073	0	0	0
24	Julie Cairney	Zongwen Liu	1.066087232	0	0	0
25	Yuan Chen	Luming Shen	1.066920931	0	0	0
26	Itai Einav	Daniel Dias Da Costa	1.068429808	0	0	0
27	Qing Li	Daniel Dias Da Costa	1.068629658	0	0	0
28	Simon Ringer	Jun Huang	1.072038616	0	0	0
29	Xiangyuan Carl Cui	Jun Huang	1.079721156	0	0	0
30	Qing Li	Itai Einav	1.081454855	0	0	0

Table 4.2 DCG and IDCG of the Hierarchical Word Mover's Distance for 681 journals

4.3 Results

The experiment is run 14 times with incrementing sample sizes, starting from 90 journals all the way to 2,622. The MAP and NDCG are two criteria to evaluate the model performance. Furthermore, the runtime performance is also examined.

Size	Hierarchical WMD	Hierarchical WMD Variant (40 topics)	Cosine Similarity (TF)	Cosine Similarity (TF–IDF)
90	1	0.630929754	0.315464877	0.315464877
171	0.919720789	0.240673214	0.388958094	0.457494526
341	0.832373073	0.193137831	0.319786282	0.340003474
511	0.815512999	0.444624315	0.565932979	0.62738678
681	0.980518393	0.478893227	0.555826683	0.571264351
851	0.939241814	0.568927496	0.740207709	0.77709615
1021	0.965944969	0.48417155	0.814583483	0.881922846
1191	0.958808222	0.7049149	0.840281614	0.866791448
1430	0.95185734	0.754000405	0.871344677	0.879576635
1593	0.922646104	0.934590826	0.8673925	0.911217237
1741	0.887107334	0.676153068	0.789503215	0.813613514
1856	0.887833555	0.796304732	0.824963342	0.889259864
1940	0.903962645	0.779194546	0.894810743	0.909369978
2622	0.969286012	0.939292887	0.94178347	0.975816595

Table 4.3 NDCG for all the models

The results as to NDCG are shown in the Table 4.3 and its graphical representation in the Fig. 4.4. For any sample sizes less than 1,500, the Hierarchical Word's Mover Distance outperforms any other models. However, as the sample size increases, the performance differences become less significant. In fact, when all 2,622 journal titles are used, differences among them become almost non-existence. The Hierarchical Word's Mover Distance works consistently well sitting anywhere between 0.82 and 1 whereas Hierarchical Word's Mover Distance Variant and Cosine Similarity generally have large variances. For the Hierarchical Word's Mover Distance Variant, the Latent Dirichlet Allocation uses words to help creating topic clusters. In a small dataset, words that are made up of topic clusters are very much alike. Therefore, the performance deteriorates. The same story roughly goes for the Cosine Similarity, which is less optimal in a lower dimension [2]. Despite the Hierarchical Word's Mover Distance performs well in a small dataset, its performance becomes comparable to other models in a large dataset.

On the other hand, the results regarding AP are shown in the Table 4.4 and its graphical representation in the Fig. 4.5. When treating co-authorships as a binary, the Hierarchical Word's Mover Distance outperforms other models regardless the sample

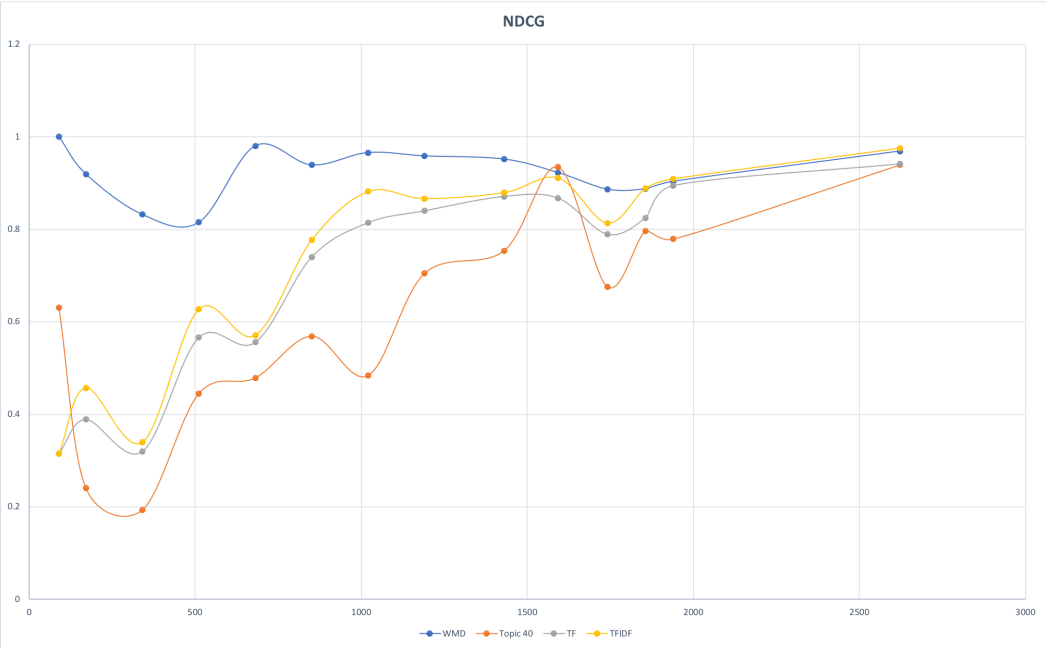

Fig. 4.4 NDCG for all the models

Size	Hierarchical WMD	Hierarchical WMD Variant (40 topics)	Cosine Similarity (TF)	Cosine Similarity (TF–IDF)
90	1	0.5	0.125	0.125
171	0.833333333	0.040151515	0.182539683	0.25
341	0.519607843	0.018531811	0.0828125	0.116666667
511	0.54622727	0.15120686	0.282992525	0.345295487
681	0.948412698	0.276701177	0.439484127	0.480494505
851	0.790223665	0.308041576	0.604619363	0.656362215
1021	0.849125429	0.35012917	0.625537379	0.662904911
1191	0.911255411	0.483768793	0.670127745	0.727506122
1430	0.802141271	0.373794095	0.662483072	0.696208318
1593	0.80882677	0.670747365	0.676497398	0.705373049
1741	0.821590725	0.460097148	0.730590793	0.764875167
1856	0.802078825	0.454028639	0.65482772	0.733915069
1940	0.820983776	0.478350593	0.756421057	0.801239405
2622	0.794309231	0.619055559	0.71473001	0.734454366
MAP	**0.803436875**	**0.370328879**	**0.514904526**	**0.557163949**

Table 4.4 AP for all the models

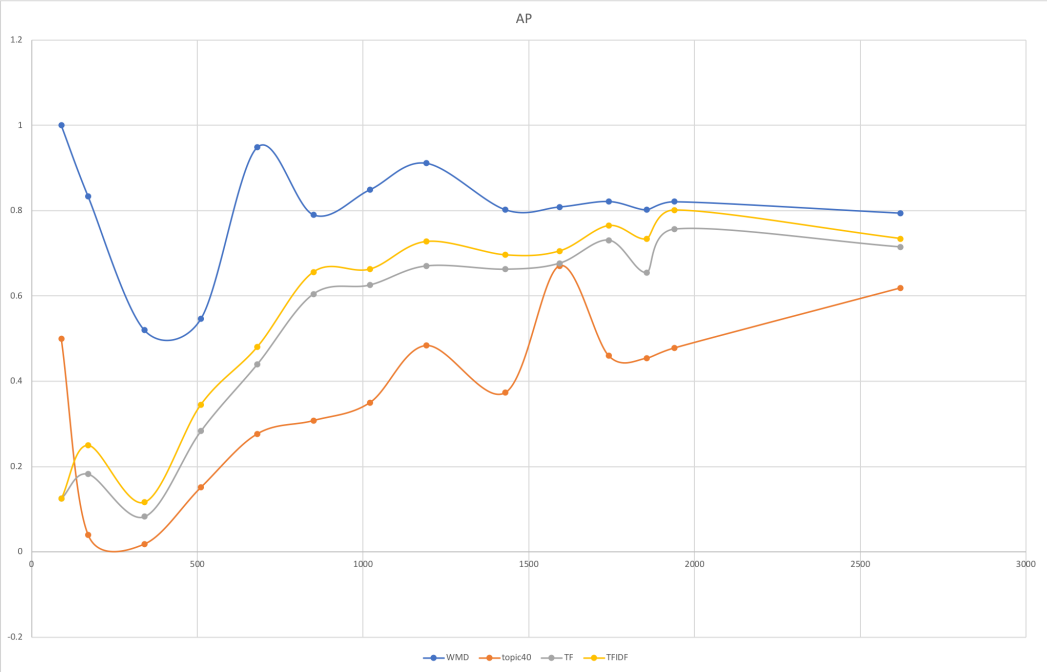

Fig. 4.5 AP for all the models

size. On the contrary, the Hierarchical Word's Mover Distance Variant and Cosine Similarity perform poorly in a small dataset. Furthermore, the MAP indicates the Hierarchical Word's Mover Distance is the best model overall.

A user would be frustrated if a query took too long to run. Hence, runtime performance is equally important when evaluating the models. The Table 4.5 and Fig. 4.6 show the runtime performance across all the models. In this category, there is clearly only one winner – Cosine Similarity (both TF and TF–IDF). The runtime performance in the Hierarchical Word's Mover Distance exhibits exponentially as the dataset increases because of pairwise comparisons. The spike that occurs in the size of 1430 of the Hierarchical Word's Mover Distance is a measurement error because a computer was asleep while the clock was still ticking.

4.4 Summary

Although the NDCG is more suited to our needs, it does not summarises values into one number that indicates an overall performance across all experiments. Oppositely, the MAP summarises a single number to provide an overall performance, but it does not take co-authorships into consideration. Unfortunately, both the NDCG and MAP reach somewhat slightly different conclusions. The MAP points out the Hierarchical

Size	Hierarchical WMD	Hierarchical WMD Variant (40 topics)	Cosine Similarity (TF)	Cosine Similarity (TF-IDF)
90	731.1674728	682.3748789	0.750457048	0.750457048
171	700.7139821	725.7942162	0.906848907	0.906848907
341	811.3287098	738.4913583	0.923554897	0.923554897
511	973.0745931	791.97663	1.05587101	1.05587101
681	1234.807673	730.0288088	1.087339163	1.087339163
851	1554.906604	799.1285028	1.255934	1.255934
1021	1898.00773	833.7420928	1.407093048	1.407093048
1191	2490.527794	878.7907021	1.415880203	1.415880203
1430	12548.41129	1034.748602	1.626306057	1.626306057
1593	3521.753737	1119.835782	1.603671074	1.603671074
1741	4426.688806	1188.600887	1.482850075	1.482850075
1856	4834.421373	1240.554867	1.683229923	1.683229923
1940	4320.762738	1236.131402	1.621683121	1.621683121
2622	7034.623559	1650.250783	2.056237221	2.056237221

Table 4.5 Runtime performance for all the models

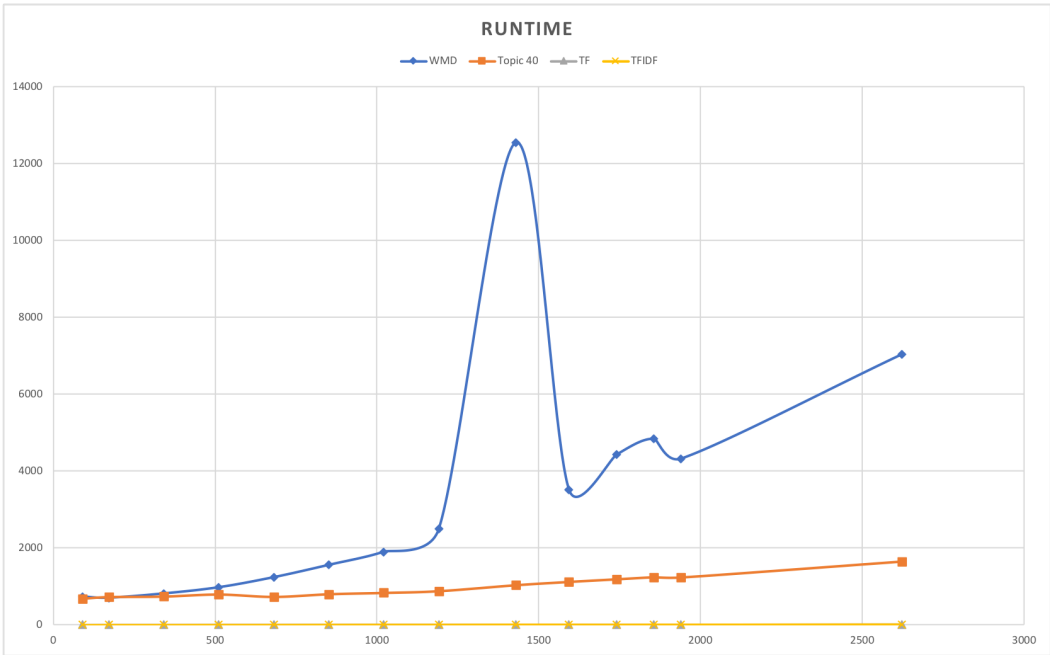

Fig. 4.6 Chart on runtime performance for all the models

Word's Mover Distance is the best model overall by some margins, but the NDCG suggests the Hierarchical Word's Mover Distance performs better than others only in a small dataset. When a sample size increases, all models are comparable. In term of runtime performance, only the Cosine Similarity wins the title. Hence, no models excel in all categories.

It is suggested a hybrid model that consists of both the Hierarchical Word's Mover Distance and either the Hierarchical Word's Mover Distance Variant or Cosine Similarity would be a better bet. In other words, for any sample sizes less than 1,500, we should run the Hierarchical Word's Mover Distance. Otherwise, the Hierarchical Word's Mover Distance Variant or Cosine Similarity should be used.

Chapter 5

Visualisations

5.1 Designs

From a user perspective, it is easier to visualise results rather than have all the distances in front of her. Hence, we have developed network visualisations. All visuals are developed using the D3.js, which is a Javascript library for creating visualisations on web pages. To further improve user experience, a web platform is created to host all visuals, which allows users to interact with them to gain further insights from the data. In this Section, we provide a few of visuals and rationales behind the designs. All D3.js visuals are using existing templates where references are provided.

5.1.1 Arc Diagram

The Arc Diagram [26], which is shown in the Fig. 5.1, visualises distances among researchers. Each researcher is drawn as a circle, and colour of circles represents a school the researcher comes from. The thickness of arcs that is connecting two researchers represents how close they are. The thicker arc is; the closer two researchers are. The thresholds dropdown has various percentage selections in which the diagram only shows the top selected percent; the order dropdown arranges circles in a either school, frequency or name sequence. At the 5% threshold, the figure clearly indicates Simon Ringer and Julie Cairney is the closest. The Table 3.2 also demonstrated this pair has the most co-authorships.

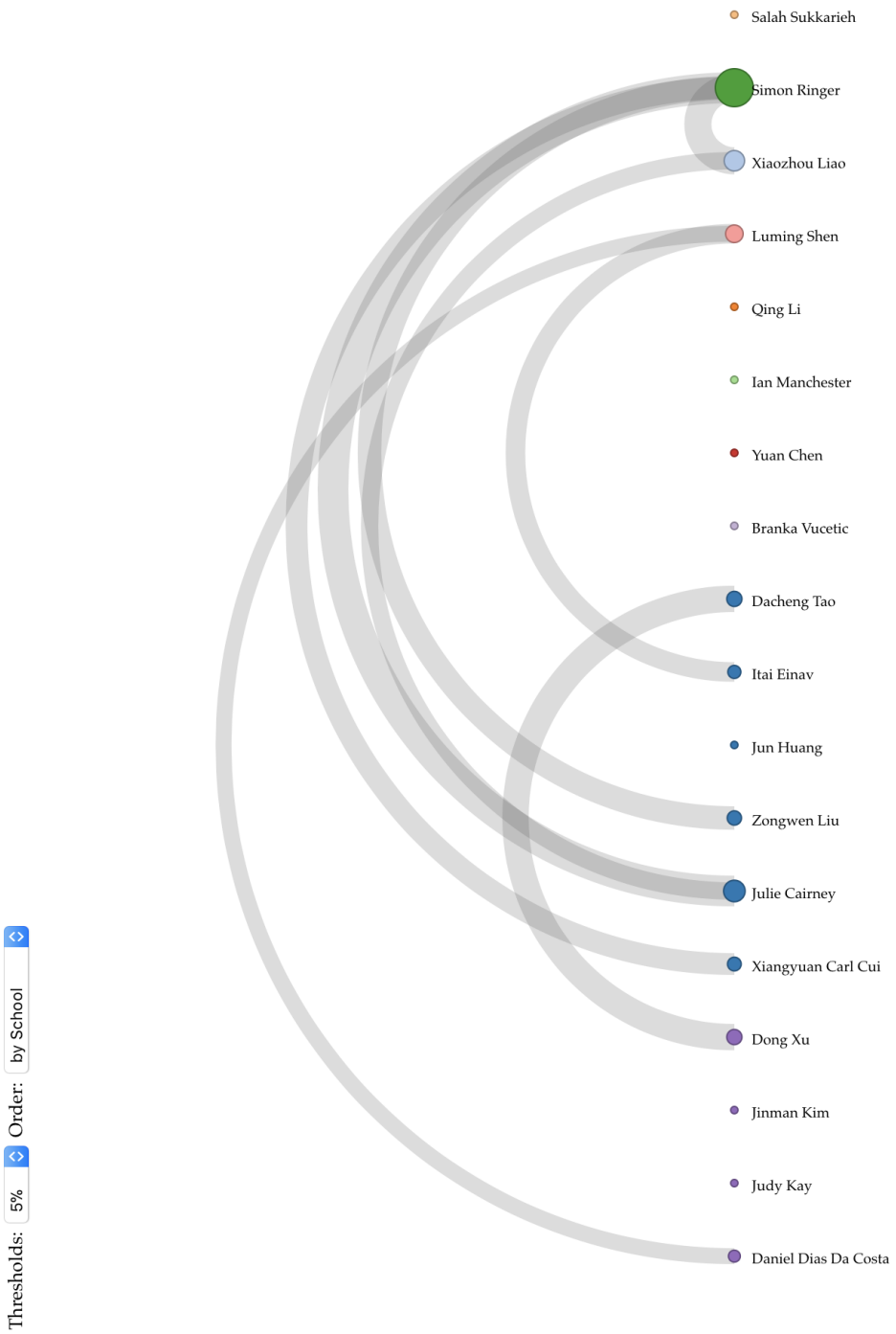

Fig. 5.1 Arc Diagram on researchers

5.1.2 Force-Directed Graph

Similar to the Arc Diagram, the Force-Directed Graph [10], which is shown in the Fig. 5.2, also visualises distances among researchers. Each researcher is drawn as a circle, and colour circles represents a school the researcher comes from. The thickness of arcs that is connecting two researchers represents how close they are. The thicker arc is; the closer two researchers are. The thresholds dropdown has various percentage selections in which the diagram only shows the top selected percent. If a mouse cursor is hovered over a circle, a corresponding researcher name will be popped up.

Fig. 5.2 Force-Directed Graph on researchers

5.1.3 Multiple Histograms

The Latent Dirichlet Allocation maps a set of words collected from journals into a mixture of a small number of topics. To examine an individual topic as well as keywords that are made up of that topic, the Multiple Histograms [19], which is shown in the Fig. 5.3, shows keywords of each topic and their corresponding weights. If a mouse cursor is hovered over a bin, a corresponding keyword will be popped up. A total number of histograms shown is a number of topics set to the Latent Dirichlet Allocation.

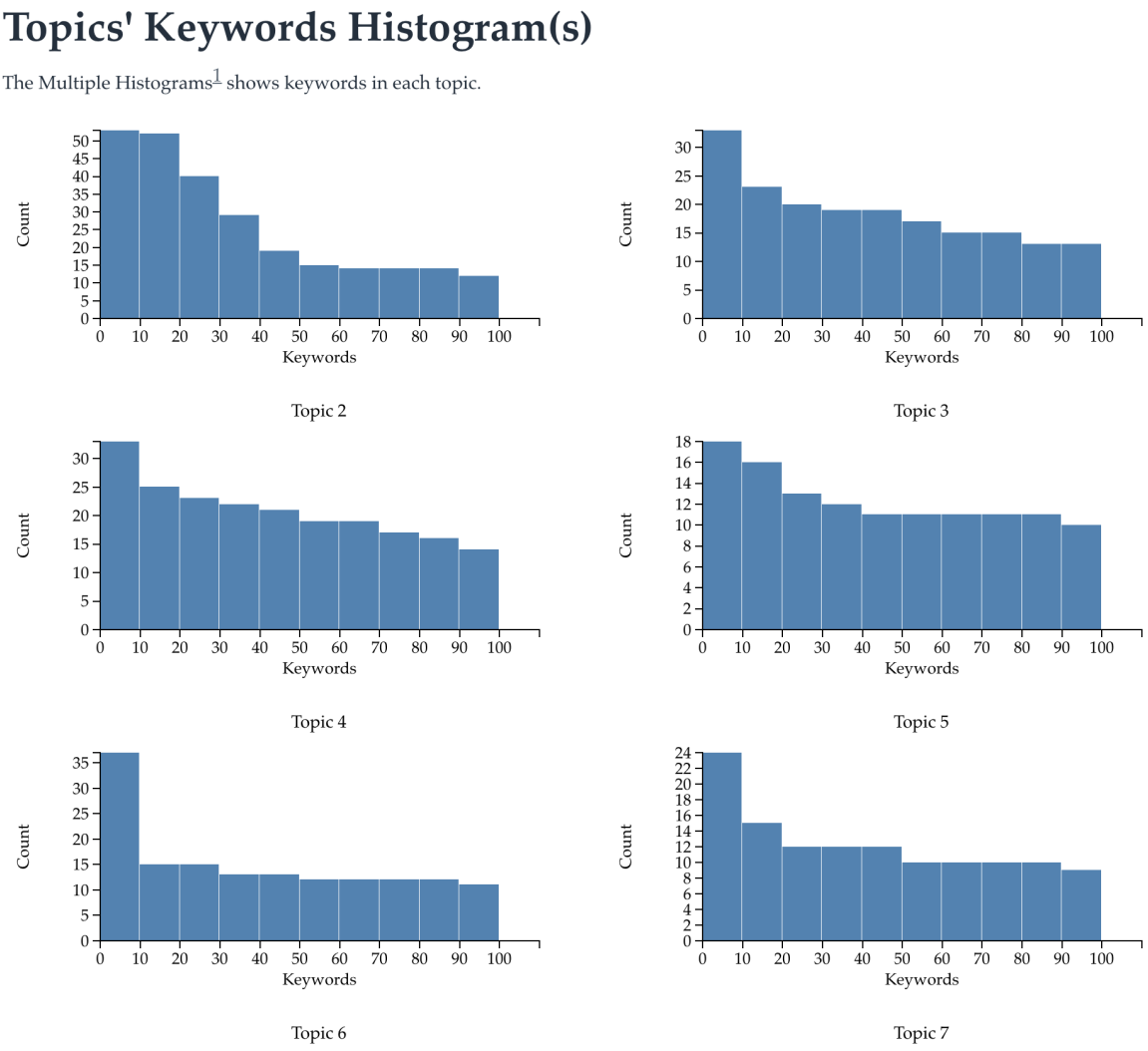

Fig. 5.3 Histograms on topics

5.1.4 Circle Packing

The Circle Packing [7], which is shown in the Fig. 5.4, visualises topics found using the Latent Dirichlet Allocation. Each bubble has researchers whose major journals fall into that topic cluster. A bigger circle is; more journals fall into that cluster. If a bubble is clicked on, it will zoom in to show their names.

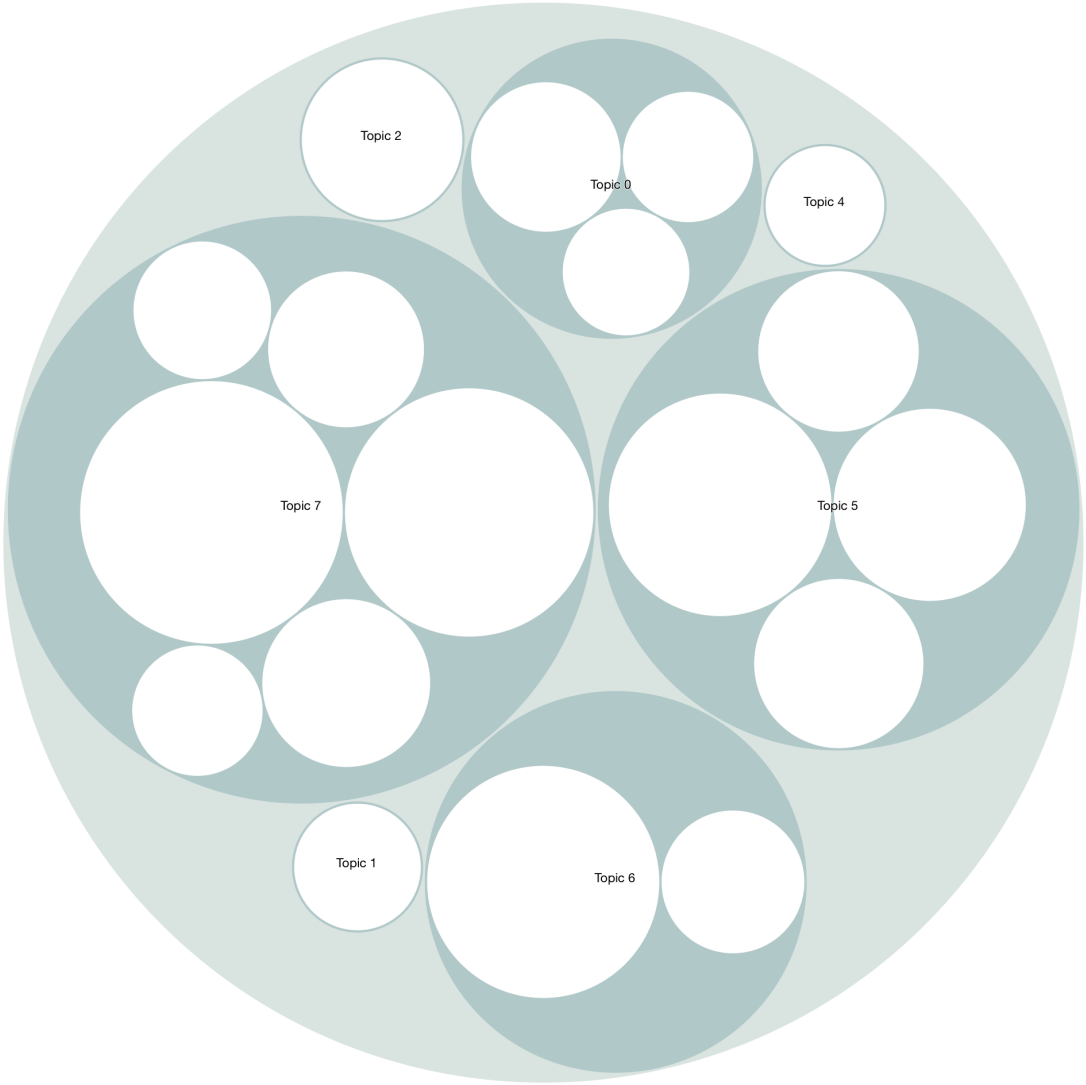

Fig. 5.4 Circle Parking on topic clusters

5.1.5 Zoomable Treemap

The Treemap diagram [16], which is shown in the Fig. 5.5, shows what interests researchers have. It is irrelevant to any parts of distance metrics, but a request from the business.

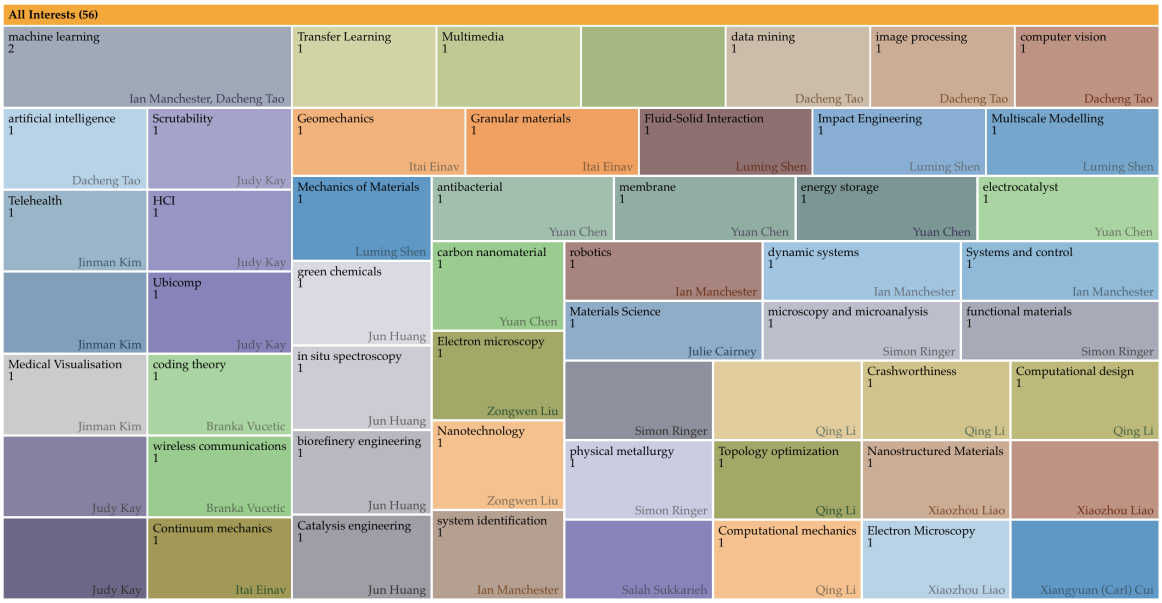

Fig. 5.5 Treemap on researcher interests

5.2 Implementation

To deploy a variety of visual tools, namely D3.js and Tableau (i.e. visuals developed by Philip), onto a unified platform is not a easy task. Eventually, we decide to host them on a web platform. It is an accessible tool since everyone has a browser (i.e. Internet Explorer, Firefox, Safari, and etc.) installed. A snapshot of the web platform is displayed in the Fig. 5.6. To embed Tableau visuals into the web platform, visuals are required to be created in the Tableau Public. Nonetheless, no additional work is required for D3.js.

The web architecture adopts the Model–View–Controller (MVC) methodology [33], which separates Javascripts, CSS sheets, and HTML pages. All Javascripts are stored in the `_script` folder; CSS sheets are stored in the `_css` folder; HTML pages are listed in the parent directory. The Model[1] manages which visuals are to show upon user

[1]The word "Model" in software engineering generally refers to the business logic whereas in the machine learning often means the statistical method. We refer to the business logic in this instance.

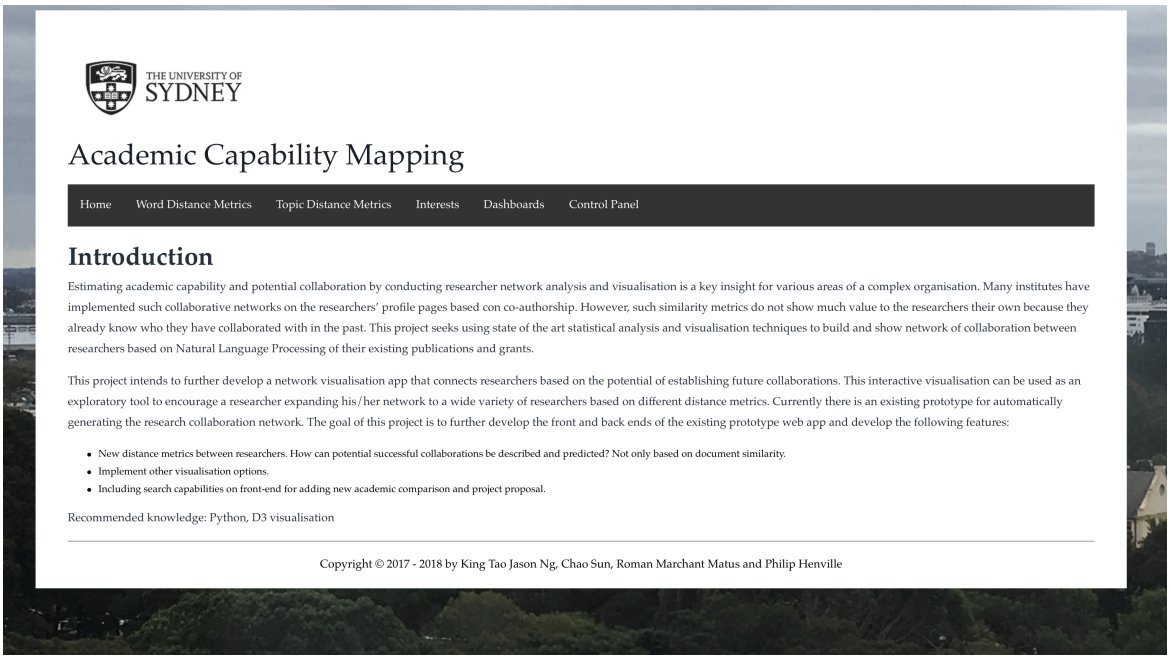

Fig. 5.6 Snapshot of the web platform

inputs. However, the distance metrics are implemented in Python. The communication between the model and distance metrics by means of `JSON` messages. In other words, the Python outputs distances in the `JSON` messages for the web application to process.

5.3 User Experience Evaluation

Similar to a model fit in machine learning, visuals also have to undergo an user experience evaluation. We mainly measure if a visual fits for purpose, and if so, how usable it is. The evaluation method consists of a set of questions in a designed survey. We now provide a summary of what feedback we have gathered from the survey.

The Arc Diagram is by far the most successful compared with the Force-Directed Graph. By selecting "by Frequency" in the Order dropdown, it is easy to navigate a pair with the closest distance. Each circle is clearly labelled. A drawback is, once many researchers are considered, the diagram is easy to go out of a boundary. On the other hand, the Force-Directed Graph can accommodate more circles than the Arc Diagram due to the graph layout. One negative about Force-Directed Graph is not informative. For instance, while two researchers that share the same colour indicate they are from the same school, colour themselves do not tell which school they actually come from. When a massive number of researchers is considered, both visuals suffer the same issue.

The visuals will be completely messed. To cater for this situation, we have created an additional webpage, which is shown in the Fig. 5.7, for users to download a distance metric into the CSV format.

Results

Using the Word's Mover Distance algorithm[1][2][3], the table below shows distances among any pairs of researchers. To download the table into the csv format, please click here to download.

Author 1	Author 2	Distance
Xiangyuan Carl Cui	Xiaozhou Liao	1.0362977615804256
Xiangyuan Carl Cui	Qing Li	1.0890418492272305
Xiangyuan Carl Cui	Salah Sukkarieh	1.1398573973171784
Xiangyuan Carl Cui	Simon Ringer	0.8955483517620002
Xiangyuan Carl Cui	Julie Cairney	1.0198726113185543
Xiangyuan Carl Cui	Ian Manchester	1.1330505209697932
Xiangyuan Carl Cui	Zongwen Liu	0.9895767704955007
Xiangyuan Carl Cui	Yuan Chen	1.059609479340928
Xiangyuan Carl Cui	Jun Huang	1.073098733110499
Xiangyuan Carl Cui	Luming Shen	1.056607828481496
Xiangyuan Carl Cui	Itai Einav	1.093933371981006
Xiangyuan Carl Cui	Daniel Dias Da Costa	1.1023560250240036

Fig. 5.7 Webpage for downloading a distance metric

The Multiple Histograms and Circle Packing are more relevant to the Latent Dirichlet Allocation than distance metrics themselves. The Multiple Histograms clearly shows keywords and their weights in each topic. Due to size constraints, only the top 10 keywords are shown. Even this visual can show many more histograms, a user less likely looks at them all. Similar to the Multiple Histograms, the Circle Packing shows a major topic of each researcher. It is believed researchers, who share the same topic, more likely work on the same area of interests.

Chapter 6

Generalised Linear Models

6.1 Descriptive Statistics

In this section, we dive into the data, which is extracted from researchers' profile pages. Data consists of

- Co-authorships for a pair of researchers
- Calculated distance metric from the Hierarchical Word Mover's Distance
- Binary indicating 1 if a pair shares the same title; otherwise 0
- Binary indicating 1 if a pair comes from the same department; otherwise 0
- Differences with regard to the duration of their contracts in days

We have 2,701 pairs in total. The Table 6.1 shows the first 30 pairs of researcher profiles. In the sample set, distance has a mean of 1.151 and a standard derivation of 0.0736. It is worth to mention the Title has been preprocessed. For instance, "Senior Lecturer", "Lecturer" or "Associate Lecturer" is mapped to "Lecturer". The response variable is co-authorships; the explanatory variables are the rest. The Fig. 6.1 plots the response counts against distance. It shows somewhat a negative relationship between the two. The Fig. 6.2 indicates almost no relationships at all between co-authorships and duration. The Fig. 6.3 demonstrates no relationships between co-authorships and title. However, the Fig. 6.4 reveals if a pair comes from the same department, it will be likely to have more co-authorships.

Co-authorships	Researcher1	Researcher2	Distance	Title	Department	Duration
25	Joseph Lizier	Mikhail Prokopenko	0.344081887	0	1	84
15	James Ward	Stewart Worrall	0.506936579	1	1	1290
31	Eduardo Nebot	Stewart Worrall	0.652820559	0	1	3279
41	Steven Armfield	Michael Kirkpatrick	0.667356028	1	1	2495
8	Albert Zomaya	Mikhail Prokopenko	0.675607872	1	0	4606
9	Joseph Lizier	Albert Zomaya	0.701469131	0	0	4690
3	Hala Zreiqat	Zufu Lu	0.713546647	1	1	1282
44	Bob Kummerfeld	Judy Kay	0.745227537	1	1	4922
15	Michael Kirkpatrick	Nicholas Williamson	0.745956446	0	1	434
24	Steven Armfield	Nicholas Williamson	0.74975835	0	1	2929
29	Craig Jin	Alistair McEwan	0.768221345	0	1	3171
4	James Ward	Asher Bender	0.785957016	1	1	747
24	Kim Rasmussen	Hao Zhang	0.791505181	0	1	3230
14	Kalina Yacef	Irena Koprinska	0.814178807	1	1	300
16	Gwenaelle Proust	Luming Shen	0.825873767	1	1	182
16	Eduardo Nebot	Hugh Durrant-Whyte	0.828363592	1	0	0
2	Hala Zreiqat	Colin Dunstan	0.829718071	0	1	736
4	Pierre Rognon	Benjy Marks	0.830670631	1	1	729
16	Hugh Durrant-Whyte	Salah Sukkarieh	0.833719254	1	0	21
26	Itai Einav	Pierre Rognon	0.837085708	0	1	2873
41	Branka Vucetic	Wibowo Hardjawana	0.837536625	0	1	3587
29	Judy Kay	Kalina Yacef	0.840943137	1	1	144
15	Eduardo Nebot	James Ward	0.850871566	0	1	4569
1	Colin Dunstan	Zufu Lu	0.852772841	0	1	546
21	Hugh Durrant-Whyte	Fabio Ramos	0.85367489	1	0	2719
9	Abbas El-Zein	Federico Maggi	0.877569836	0	1	3582
4	Stewart Worrall	Asher Bender	0.887522771	1	1	2037
15	Itai Einav	Benjy Marks	0.891481245	0	1	3602
3	Fernando Alonso-Marroquin	Yixiang Gan	0.903393	1	1	197
18	Branka Vucetic	Mahyar Shirvanimoghaddam	0.924793515	0	1	6265

Table 6.1 Sample data of a pair of researcher profiles

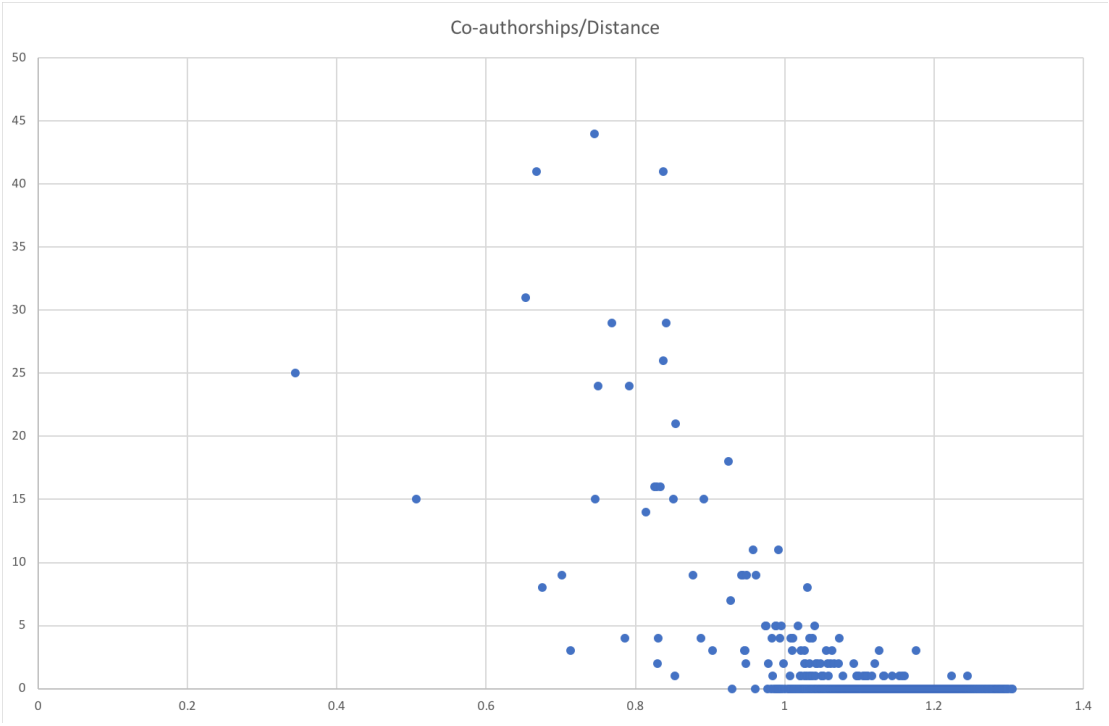

Fig. 6.1 Relationship between co-authorships and distance

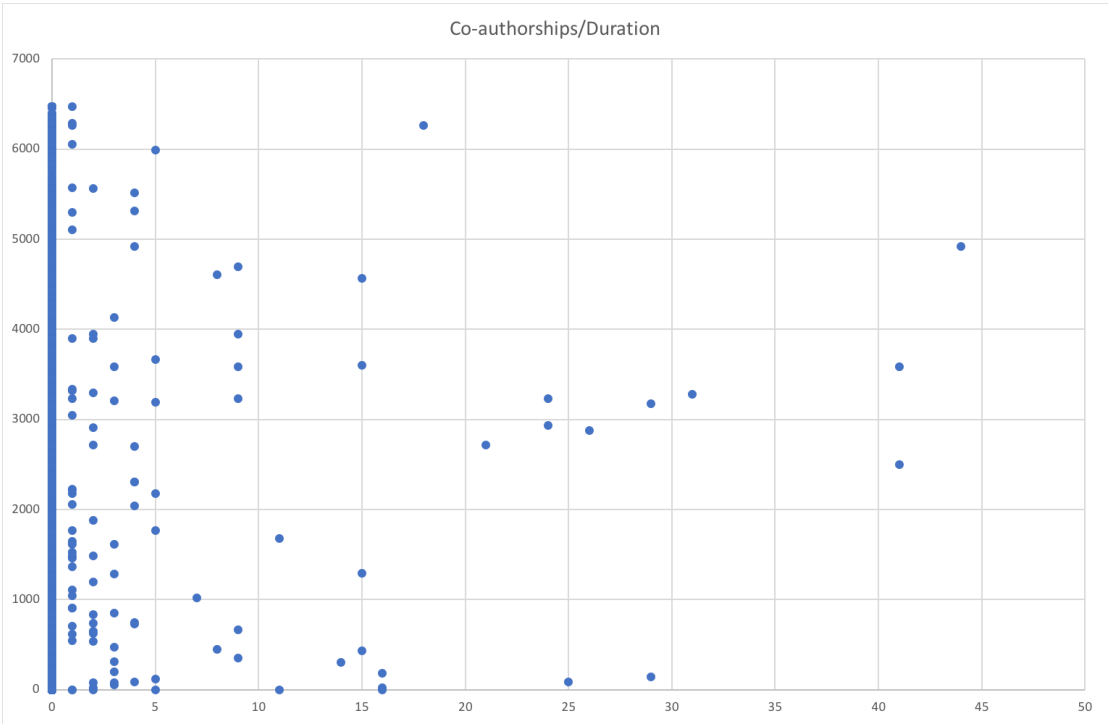

Fig. 6.2 Relationship between co-authorships and duration

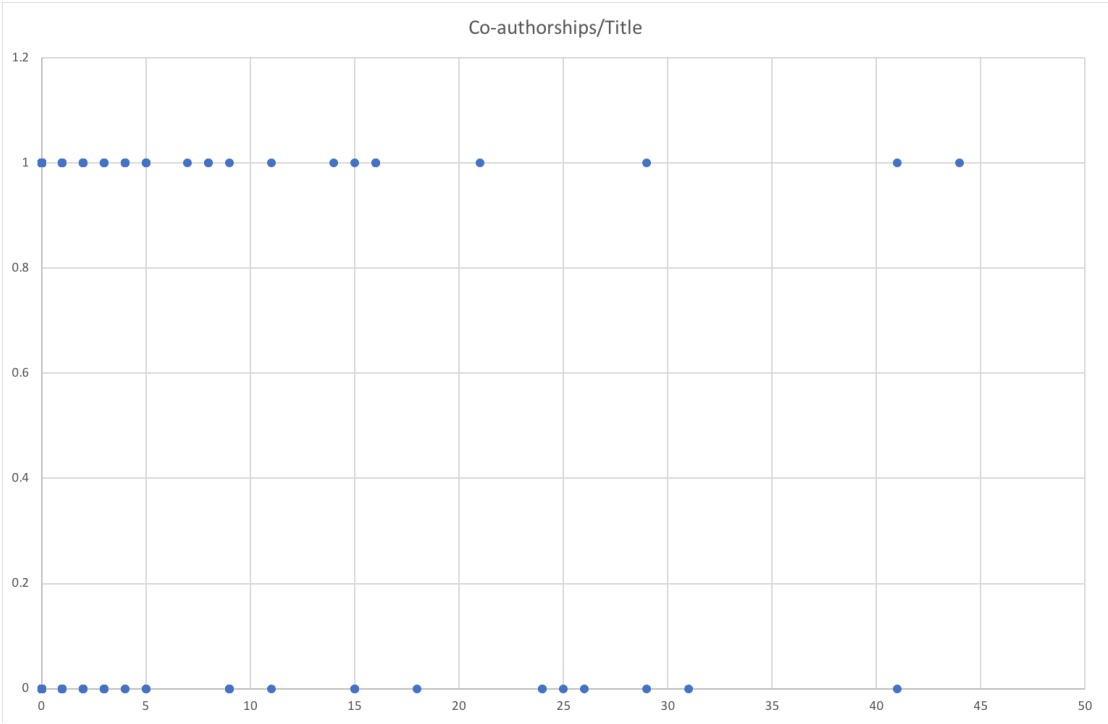

Fig. 6.3 Relationship between co-authorships and title

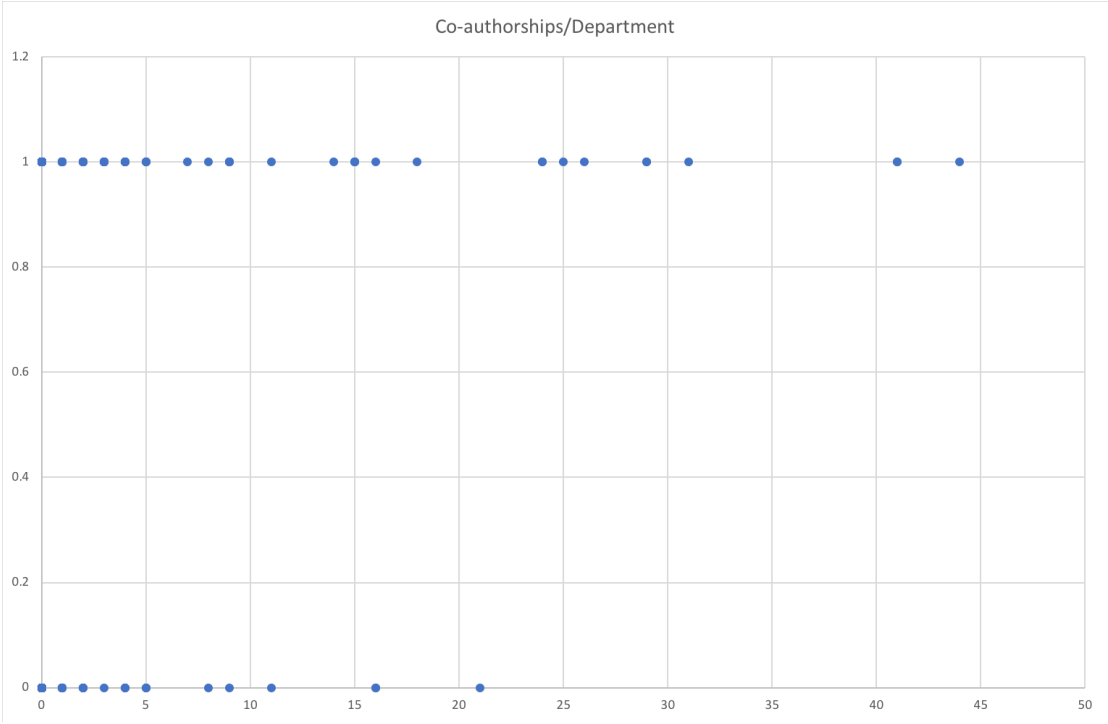

Fig. 6.4 Relationship between co-authorships and department

6.2 Negative Binomial Regression

While publications are a critical factor for researcher recommendations, social factors are equally important. Using a calculated distance metric from the Hierarchical Word Mover's Distance as a predictor variable along with duration of contracts, professional titles, and departments, we use a *Generalized Linear Model* (GLM) to determine whether any predictor variables affect co-authorships.

All GLMs have three components – *Random Component*, *Systematic Component* and *Link Function* [4]. We assume a random component follows the Poisson distribution as co-authorships are count data. Poisson distributions require both mean and variance must be identical. Unfortunately, in our case, the variances are larger than the means. In statistics, this phenomenon is called *Overdispersion* [4]. Using the overdispersion test [12], there is evidence of overdispersion because it fails to reject the null hypothesis at the 0.05 significance level. On the other hand, the Negative Binomial distribution allows the variance exceeds the mean as it has an additional parameter, called *dispersion parameter*, to capture the variance [4]. We assume it follows the Cauchy distribution with a location parameter, $x_0 = 0.015$ and a scale parameter, $\gamma = 0.015$. Therefore, we propose the Negative Binomial GLM

$$\log(y) = \beta_0 - \beta_1 x_{distance} + \beta_2 x_{duration} + \beta_3 x_{title} + \beta_4 x_{department} \tag{6.1}$$

We then fit it using the Bayesian statistics [21]

$$\log(\hat{y}) = 22.83 - 24.02 \times x_{distance} + 0.000033 \times x_{duration} + 0.23 \times x_{title} + 1.34 \times x_{department}, \tag{6.2}$$

where $\hat{y}$ is the expected co-authorship counts, $x_{distance}$ is a calculated distance metric from the Hierarchical Word Mover's Distance between two researchers, $x_{duration}$ is the difference with regard to the duration of their contracts in days, x_{title} is a binary indicating 1 if a pair shares the same title, and $x_{department}$ is a binary indicating 1 if a pair comes from the same department.

The Table 6.2 and Fig. 6.5 show the posterior distributions for all coefficients (i.e. ϕ is the Cauchy distribution) from PyStan; the Table 6.3 and Fig. 6.6 show the same distributions from PyStan with normalized data using $z_{ij} = \frac{x_{ij} - \mu_i}{\sigma i}$, where $i \in (distance, duration, title, department)$ and j is the jth pair of researchers. Since $\hat{\beta}_2 \approx 0$ and $\hat{\beta}_3 \approx 0$, they are less signification compared with $\hat{\beta}_1$ and $\hat{\beta}_4$. The model indicates a shorter distance yields more co-authorships. In other words, it is suggested a pair, who individually writes similar journals, yields more co-authorships. The model

	mean	**se_mean**	**sd**	**2.50%**	**25%**	**50%**	**75%**	**97.50%**	**n_eff**	**Rhat**
ϕ	0.26	1.00E-03	0.04	0.18	0.23	0.25	0.28	0.35	1797	1
β_0	22.83	0.05	1.81	19.43	21.63	22.78	23.97	26.68	1149	1
β_1	-24.02	0.05	1.68	-27.51	-25.07	-23.96	-22.9	-20.76	1202	1
β_2	3.30E-05	1.10E-06	6.10E-05	-8.90E-05	-7.60E-06	3.10E-05	7.40E-05	1.50E-04	2867	1
β_3	0.23	5.80E-03	0.24	-0.23	0.07	0.23	0.39	0.69	1688	1
β_4	1.34	5.20E-03	0.23	0.89	1.19	1.34	1.5	1.79	1882	1
lp___	826.26	6.00E-02	1.82	821.75	825.32	826.61	827.6	828.71	1055	1

Table 6.2 Posterior distributions of all coefficients

	mean	**se_mean**	**sd**	**2.50%**	**25%**	**50%**	**75%**	**97.50%**	**n_eff**	**Rhat**
ϕ	0.34	1.60E-03	0.08	0.22	0.29	0.33	0.39	0.51	2294	1
β_0	-6.02	8.20E-03	0.39	-6.86	-6.26	-5.99	-5.75	-5.31	2285	1
β_1	-1.7	3.40E-03	0.15	-2.03	-1.81	-1.69	-1.59	-1.43	2119	1
β_2	-1.00E-02	3.20E-03	1.70E-01	-3.60E-01	-1.40E-01	-1.00E-02	1.00E-01	3.20E-01	2975	1
β_3	0.12	6.50E-03	0.35	-0.59	-0.11	0.13	0.37	0.81	2937	1
β_4	1.21	6.90E-03	0.36	0.52	0.97	1.21	1.45	1.89	2666	1
lp___	77.4	5.00E-02	1.84	72.84	76.38	77.75	78.73	79.89	1383	1

Table 6.3 Posterior distributions of all coefficients (Normalized data)

also concludes a pair, who comes from the same department, yields more co-authorships. We believe in this case a pair are in close proximity.

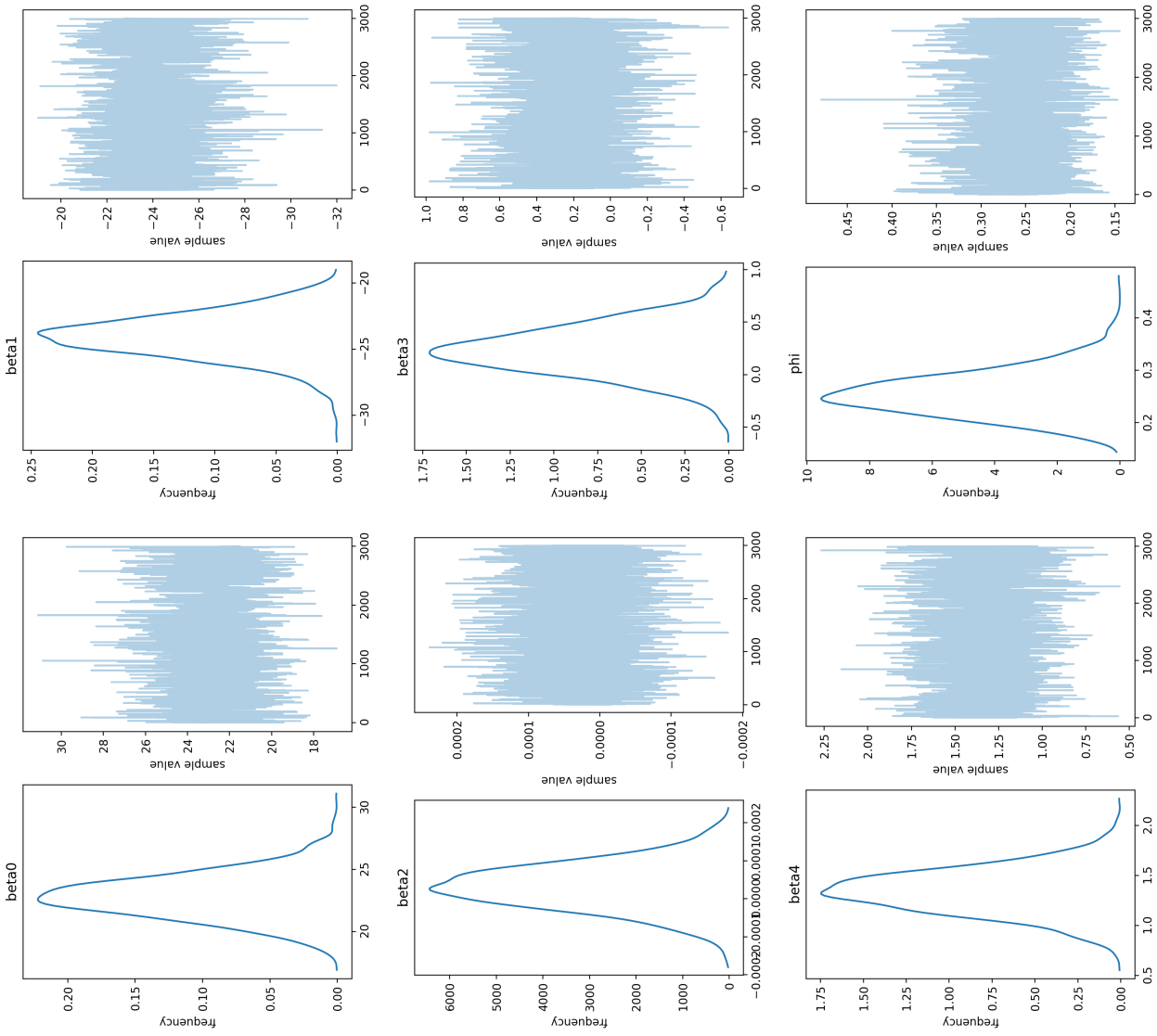

Fig. 6.5 Posterior distributions of all coefficients

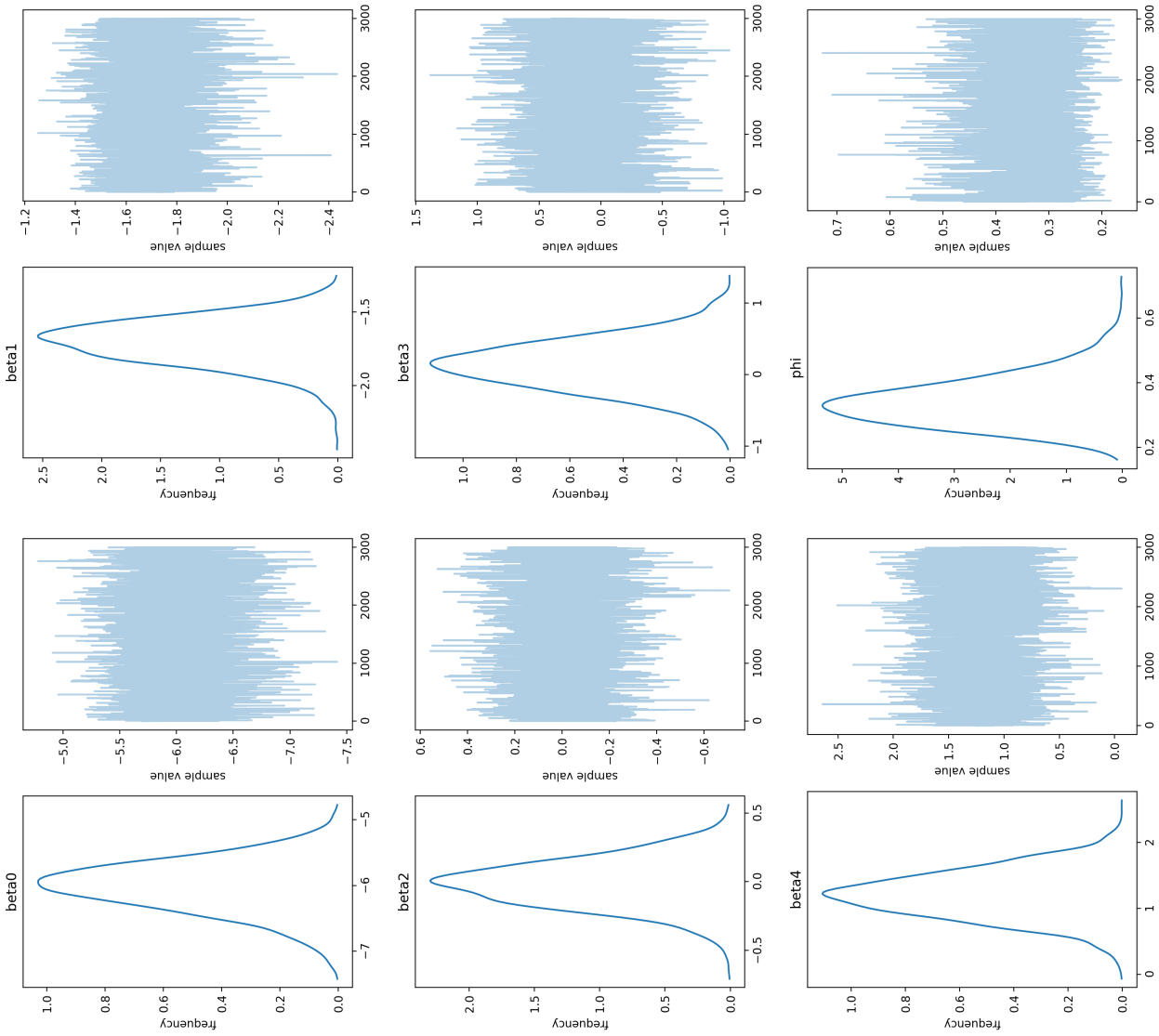

Fig. 6.6 Posterior distributions of all coefficients (Normalized data)

6.3 Generalised Linear Mixed Models

While distance and department variables are significant, it would be difficult to convince an individual department is equally significant. A more convincing argument would be one or two departments may be more significant than others. In other words, the department-specific term takes the same value for all observations in that department. Therefore, we introduce a random effect or hierarchical structure into the Negative Binomial GLM, or more precisely Generalised Linear Mixed Models (GLMM) [4]. Provided a pair from the same department, we use a GLMM to determine if an individual department is more significant than others.

$$\log(y_{ij}) = u_i - \beta_i x_{ij}, \tag{6.3}$$

where y is the co-authorship counts, i is the ith department, and j is the jth pair of researchers. Each department now has its own intercept term, u_i, and distance term, β_i. We have the following departments — Aerospace Mechanical & Mechatronic Engineering, Australian Centre for Field Robotics, Chemical & Biomolecular Engineering, Civil Engineering, Electrical & Information Engineering, and Information Technologies.

We fit the model 6.3 using Bayesian statistics [21]. As we can see the partial result in the Table 6.4 and Fig. 6.7 from PyStan, u_3 and u_5 are more important than the rest; β_3 and β_5 are also more important than the others. In other words, if a pair comes from the same department, and the department is either Australian Centre for Field Robotics or Electrical & Information Engineering, it will be likely to have more co-authorships. For the sake of completeness, μ_u is the mean of u; σ_u, which is the variance of u, is assumed to follow the uniform distribution.

	mean	**se_mean**	**sd**	**2.50%**	**25%**	**50%**	**75%**	**97.50%**	**n_eff**	**Rhat**
u_1	14.32	0.04	2.17	10.57	12.78	14.16	15.67	18.96	3264	1
u_2	13.04	0.06	3.4	7.25	10.57	12.76	15.27	20.24	3036	1
u_3	26.85	0.17	7.37	15.66	21.35	25.82	31.15	43.81	1957	1
u_4	19.85	0.03	2.08	16.1	18.44	19.75	21.19	24.35	3929	1
u_5	25.13	0.13	6.1	15.78	20.67	24.25	28.75	39.01	2311	1
u_6	22	0.07	3.83	15.57	19.32	21.62	24.22	30.49	2819	1
ϕ	0.58	1.70E-03	0.12	0.38	0.49	0.56	0.65	0.85	4794	1
β_1	-15.38	0.04	2.15	-19.94	-16.73	-15.24	-13.86	-11.57	3283	1
β_2	-12.87	0.06	3.42	-20.18	-15.13	-12.59	-10.4	-6.92	3047	1
β_3	-25.81	0.16	7.02	-41.89	-29.93	-24.88	-20.6	-15.08	1956	1
β_4	-19.7	0.03	2.01	-24.05	-20.98	-19.6	-18.33	-16.08	3917	1
β_5	-25.39	0.13	6.47	-40.14	-29.21	-24.46	-20.64	-15.58	2304	1
β_6	-21.34	0.07	3.57	-29.26	-23.38	-21	-18.85	-15.29	2819	1
μ_u	20.37	0.1	5.13	12.07	17.46	19.83	22.68	32.17	2449	1
σ_u	9.13	0.17	6.28	1.89	5.18	7.69	11.36	25.53	1388	1
$\hat{y_1}$	13.07	0.02	1.4	10.57	12.13	12.99	13.97	16.09	3943	1
$\hat{y_2}$	6.52	0.03	1.68	3.68	5.31	6.37	7.61	10.08	3043	1
$\hat{y_3}$	4.64	0.02	1.19	2.64	3.77	4.54	5.41	7.21	3071	1
$\hat{y_4}$	4.06	0.01	0.79	2.73	3.51	3.99	4.53	5.77	3438	1
$\hat{y_5}$	3.35	0.01	0.7	2.17	2.86	3.28	3.76	4.9	3511	1
$\hat{y_6}$	6.09	0.02	1.2	4.1	5.24	5.97	6.81	8.76	2882	1
$\hat{y_7}$	2.85	0.01	0.64	1.78	2.41	2.79	3.23	4.28	3584	1
$\hat{y_8}$	2.79	0.01	0.63	1.73	2.36	2.73	3.17	4.2	3595	1
$\hat{y_9}$	5.63	0.02	1.24	3.63	4.73	5.49	6.38	8.49	2625	1
$\hat{y_{10}}$	2.93	0.01	0.77	1.64	2.38	2.86	3.41	4.61	3196	1
$\hat{y_{11}}$	4.26	8.10E-03	0.52	3.32	3.9	4.23	4.6	5.38	4109	1
$\hat{y_{12}}$	4.62	0.02	0.96	3.04	3.94	4.52	5.2	6.76	2930	1
$\hat{y_{13}}$	3.58	7.10E-03	0.46	2.75	3.27	3.55	3.88	4.55	4169	1
$\hat{y_{14}}$	1.56	8.00E-03	0.5	0.71	1.21	1.53	1.86	2.68	3937	1
$\hat{y_{15}}$	3.49	6.90E-03	0.45	2.67	3.18	3.46	3.78	4.43	4180	1
$\hat{y_{16}}$	3.36	6.70E-03	0.44	2.57	3.06	3.34	3.65	4.28	4195	1
$\hat{y_{17}}$	3.87	0.02	0.87	2.43	3.25	3.79	4.39	5.87	3121	1
$\hat{y_{18}}$	4.05	0.02	0.87	2.61	3.43	3.96	4.57	5.98	2960	1
$\hat{y_{19}}$	2.09	9.80E-03	0.57	1.12	1.69	2.04	2.46	3.33	3422	1
$\hat{y_{20}}$	1.21	7.30E-03	0.47	0.4	0.89	1.18	1.49	2.24	4105	1
$\hat{y_{21}}$	2.56	5.50E-03	0.36	1.89	2.31	2.55	2.8	3.33	4315	1
$\hat{y_{22}}$	1.62	7.80E-03	0.48	0.8	1.28	1.59	1.93	2.66	3718	1
$\hat{y_{23}}$	2.29	5.10E-03	0.34	1.67	2.06	2.27	2.51	3	4374	1
$\hat{y_{24}}$	2.06	4.80E-03	0.32	1.47	1.83	2.04	2.26	2.73	4436	1
$\hat{y_{25}}$	1.65	7.60E-03	0.59	0.6	1.26	1.63	2.03	2.9	5974	1
$\hat{y_{26}}$	1.57	4.10E-03	0.28	1.06	1.37	1.56	1.76	2.16	4607	1
$\hat{y_{27}}$	0.02	5.30E-03	0.38	-0.68	-0.23	0.01	0.26	0.83	5047	1
$\hat{y_{28}}$	1.29	3.80E-03	0.26	0.81	1.11	1.28	1.46	1.84	4742	1
$\hat{y_{29}}$	2.47	0.02	0.85	1.09	1.87	2.36	2.94	4.46	2405	1
$\hat{y_{30}}$	1.22	3.70E-03	0.26	0.75	1.04	1.21	1.39	1.76	4780	1
$\hat{y_{31}}$	1.19	3.70E-03	0.25	0.72	1.01	1.18	1.35	1.72	4799	1
$\hat{y_{32}}$	0.85	5.10E-03	0.36	0.18	0.6	0.84	1.08	1.6	4902	1
$\hat{y_{33}}$	1.16	3.60E-03	0.25	0.69	0.98	1.15	1.32	1.68	4819	1
$\hat{y_{34}}$	0.99	3.40E-03	0.24	0.54	0.83	0.98	1.15	1.49	4922	1
$\hat{y_{35}}$	-0.45	4.80E-03	0.36	-1.13	-0.69	-0.45	-0.21	0.29	5651	1
$\hat{y_{36}}$	0.92	3.30E-03	0.23	0.48	0.75	0.91	1.07	1.41	4975	1
$\hat{y_{37}}$	0.51	4.50E-03	0.34	-0.14	0.28	0.5	0.72	1.19	5692	1
$\hat{y_{38}}$	0.58	3.00E-03	0.22	0.17	0.43	0.58	0.72	1.02	5240	1
$\hat{y_{39}}$	0.4	4.40E-03	0.33	-0.24	0.18	0.39	0.61	1.08	5861	1
$\hat{y_{40}}$	1.34	0.01	0.61	0.32	0.92	1.28	1.69	2.76	3192	1
$\hat{y_{41}}$	1.15	9.80E-03	0.58	0.17	0.75	1.09	1.48	2.47	3412	1
$\hat{y_{42}}$	-1.03	4.50E-03	0.35	-1.74	-1.27	-1.02	-0.8	-0.37	5999	1
$\hat{y_{43}}$	0.11	4.40E-03	0.34	-0.55	-0.11	0.11	0.33	0.79	6004	1
$\hat{y_{44}}$	0.08	4.40E-03	0.34	-0.58	-0.14	0.08	0.31	0.76	5990	1
lp___	843.05	0.1	3.41	835.36	841.02	843.39	845.44	848.65	1232	1

Table 6.4 Posterior distributions of all coefficients for GLMM

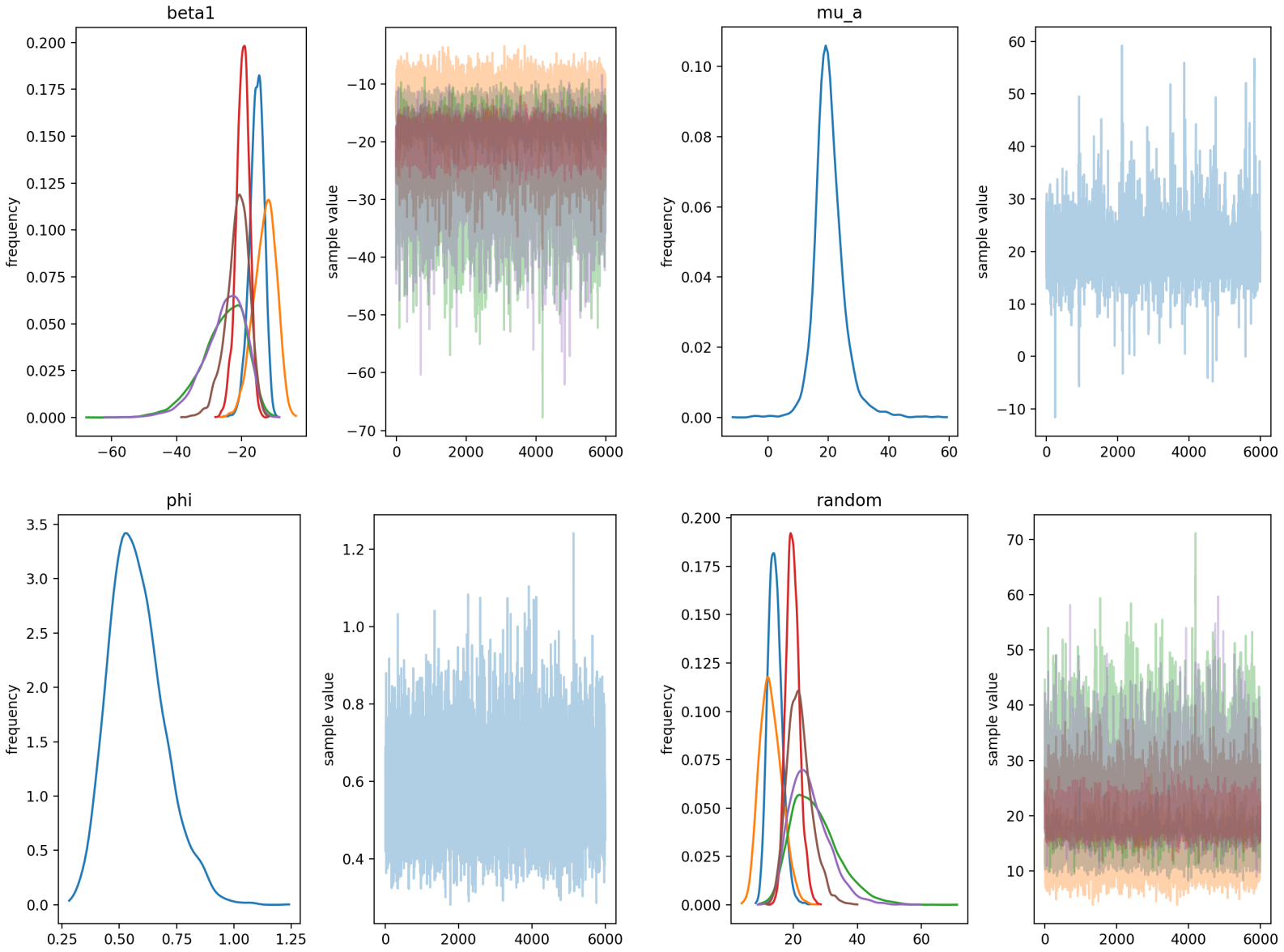

Fig. 6.7 Posterior distributions of all coefficients for GLMM

Chapter 7

Conclusion

7.1 Concluding Remarks

We previously showed the Hierarchical Word Mover's Distance outperforms other models only in a small dataset. Hence, we recommended a hybrid model that consists of both the Hierarchical Word's Mover Distance, and either its variant or Cosine Similarity. In this chapter, we like to re-visit some of remarks.

Firstly, we conducted the experiments on titles without giving any explanations. In fact, the original experimental set-up was to take abstracts rather than titles. When we first ran experiments on abstracts, we realised the Cosine Similarity performed well even in a small dataset. However, when shrinking a number of words down to have a title, the Cosine Similarity deteriorated the performance. It is argued two titles less likely share the same words compared with two abstracts. Oppositely, the Hierarchical Word's Mover Distance still performed exceptionally well in this situation. Thankfully, the word2vec model was able to work out a distance between two distinct words. It is believed this prior knowledge that helped the Word's Mover Distance to excel.

Secondly, the Hierarchical Word's Mover Distance Variant has a hyperparameter. That is, a number of topics. It is the parameter that the Latent Dirichlet Allocation requires to find topic clusters. Hence, the Hierarchical Word's Mover Distance Variant is not strictly considered as an unsupervised learning method. A number of topics will be likely different depending on what an sample is chosen. Fortunately, provided a huge dataset, performance becomes comparable when a large number of topics is set.

Finally, we mentioned data was extracted from both Google Scholarly and Scopus. Because there is a throttle issue, or rather a limit imposed from Google on how many journals can be downloaded a day, it is suggested to use Scopus to extract journals in a production environment.

In the GLM, we concluded the model indicated a shorter distance would yield more co-authorships; the model also concluded a pair from the same department would yield more co-authorships. The magnitudes are very comparable to ones generated from the maximum likelihood approach mainly because of uninformative prior being used in the Bayesian setting. Noticeably, the runtime is slower in the Markov chain Monte Carlo than Newton Raphson method. In the GLMM, the model suggested if a pair came from the same department, and the department was either Australian Centre for Field Robotics or Electrical & Information Engineering, it would have more co-authorships. Unfortunately, the maximum likelihood approach did not converge, so results could not be compared against.

7.2 Future Works

We briefly discussed Google Scholarly and Scopus in the previous section. While Scopus does address the throttle issue, it does not allow journals to be downloaded outside of the university network. Unfortunately, the remote access from home does not fix this limitation either. We also encountered data quality issues. For instance, a few journals downloaded have unprintable characters on titles to cause a flow on effect to the Elasticsearch, and they have since been removed from Scopus. Hence, it is quite possible to have created selection bias in the dataset. A more sensible approach would be to remove unprintable characters from titles.

As discussed in the Section 2.6, the runtime performance is a disadvantage of the Word's Mover Distance. While the Hierarchical Word's Mover Distance Variant attempts to address this drawback, it foregoes the performance. Brokos et al. [11] recommended to use weighted centroids of word embeddings to work out the top-k documents, so the Word's Mover Distance would only process the top-k documents. Furthermore, G. Huang et al. [18] proposed the Supervised Word Mover's Distance in which co-authorships would also be taken into account in the training. Due to time constraints, we have not implemented these methods in our experiments.

Regarding visualisations, only two visuals have been implemented — Arc Diagram and Force-Directed Graph. The Chord Diagram [8] and Dendrogram [9] could also be implemented. Data in the Chord Diagram is arranged radially around a circle, so it less likely takes up a lot of spaces compared with Arc Diagram and Force-Directed Graph; the Dendrogram is able to demonstrate a hierarchical structure. Clearly, the Arc Diagram and Force-Directed Graph are not. We also introduced Node.js, Angularjs

and Elasticsearch into the software in which some are part of the MEAN stack [15]. Due to time constraints, we have not explored its full potential.

In the Section 5.2, we did not walk through the detailed implementation. Indeed, a lot of design principles that should be in the thesis have been left out, such as the Unified Modelling Language. However, we would definitely include it if the thesis was a software development project. On the other hand, the original implementation was completely procedural. As the software grows bigger, it becomes less flexible. Scalability is a driving factor to move toward the object-oriented methodology. While we adopted a few software design principles, such as Facade Pattern, other design patterns, such as Singleton, were relatively difficult to implement in Python. Admittedly, we feel more comfortable with C++. A good reference about the Design Patterns is [33].

To incorporate social factors into the GLM is not a difficult task to a person, who once studied econometrics. However, what becomes difficult is to use the Bayesian estimation to estimate coefficients. It is certainly a challenging task. However, when we moved from the GLM to GLMM, we concluded that given a pair from the same department, Australian Centre for Field Robotics or Electrical & Information Engineering was more significant than other departments. Actually, in the first attempt, we presented the model without a varying distance term, but we realised an intercept for a particular department was similar to ones in other departments. We eventually disregarded it. Additionally, we have not implemented the predicative distribution for the GLM and GLMM due to time constraints again.

References

[1] Wikipedia - discounted cumulative gain, 2018. URL https://en.wikipedia.org/wiki/Discounted_cumulative_gain.

[2] Stackexchange - drawbacks with cosine similarity, 2018. URL https://stats.stackexchange.com/questions/266979/drawbacks-with-cosine-similarity.

[3] Wikipedia - latent dirichlet allocation, 2018. URL https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation.

[4] A. Agresti. *An Introduction to Categorical Data Analysis*. John Wiley & Sons, 2006.

[5] A. Bakharia. Using tsne to plot a subset of similar words from word2vec, November 2017. URL https://medium.com/@aneesha/using-tsne-to-plot-a-subset-of-similar-words-from-word2vec-bb8eeaea6229.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003. URL http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

[7] M. Bostock. D3 circle packing - observable. 2017. URL https://beta.observablehq.com/@mbostock/d3-circle-packing.

[8] M. Bostock. D3 chord dependency diagram. 2017. URL https://beta.observablehq.com/@mbostock/d3-chord-dependency-diagram.

[9] M. Bostock. D3 cluster dendrogram. 2017. URL https://beta.observablehq.com/@mbostock/d3-cluster-dendrogram.

[10] M. Bostock. Force-directed graph - observable. 2017. URL https://beta.observablehq.com/@mbostock/d3-force-directed-graph.

[11] G.-I. Brokos, P. Malakasiotis, and I. Androutsopoulos. Using centroids of word embeddings and word mover's distance for biomedical document retrieval in question answering. *BioNLP*, 2016. URL https://arxiv.org/abs/1608.03905.

[12] A. Cameron and P. K.Trivedi. Regression-based tests for overdispersion in the poisson model. 1990. URL https://www.sciencedirect.com/science/article/pii/030440769090014K.

[13] B. Chalmers. Scholarly · pypi, 2018. URL https://pypi.org/project/scholarly/.

[14] R. Cummins. Lecture 6: Evaluation information retrieval computer science tripos part ii, 2016. URL https://www.cl.cam.ac.uk/teaching/1516/InfoRtrv/lecture6-evaluation.pdf.

[15] E. Elrom. *Pro MEAN Stack Development*. Apress, Berkeley, CA, 2016.

[16] ganeshv. Zoomable treemap template - bl.ocks.org. 2018. URL http://bl.ocks.org/ganeshv/6a8e9ada3ab7f2d88022.

[17] S. D. Gollapalli, P. Mitra, and C. L. Giles. 'similar researcher search' in academic environments. *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, 2012. URL https://dl.acm.org/citation.cfm?id=2232849.

[18] G. Huang, C. Guo, M. J. Kusner, Y. Sun, K. Q. Weinberger, and F. Sha. Supervised word mover's distance. 2016. URL https://papers.nips.cc/paper/6139-supervised-word-movers-distance.pdf.

[19] jrzerr. D3.js multiple histograms with pre-computed histogram bins from json data. 2013. URL http://plnkr.co/agRZx6.

[20] J. Kitchin and M. E. Rose. Scopus: Python-based api-wrapper to access scopus, 2017. URL https://scopus.readthedocs.io/en/latest/.

[21] J. K. Kruschke. *Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan*. Academic Press / Elsevier, 2015.

[22] W. Kurt. Kullback-leibler divergence explained, May 2017. URL https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained.

[23] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger. From word embeddings to document distances. *Proceedings of the 32nd International Conference on Machine Learning*, 37, 2015. URL http://proceedings.mlr.press/v37/kusnerb15.pdf.

[24] J. Leskovec, A. Rajaraman, and J. D. Ullman. *Mining of Massive Datasets*. Cambridge University Press, Stanford University, California, 2014. URL http://infolab.stanford.edu/~ullman/mmds/book.pdf.

[25] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. URL https://nlp.stanford.edu/IR-book/html/htmledition/irbook.html.

[26] mayblue9. Arc diagram - bl.ocks.org. 2017. URL http://bl.ocks.org/mayblue9/dcc49ef6e3888f37f755177c4a248f2c.

[27] K. P. Murphy. *Machine Learning A Probabilistic Perspective*. Massachusetts Institute of Technology, 2012.

[28] M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, 101(1):5200–5205, 2004. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC387296/?tool=pmcentrez.

[29] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, The Pitt Building, Trumpington Street, Cambridge CB2 1RP, 2012.

[30] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40, 11/2000. URL https://www.cs.cmu.edu/~efros/courses/LBMV07/Papers/rubner-jcviu-00.pdf.

[31] C. Sun, K. T. Ng, R. M. Matus, and P. Henville. Hierarchical word mover distance for collaboration recommender system. 2018.

[32] Vector-Representations-of-Words. Vector representations of words, September 2018. URL https://www.tensorflow.org/tutorials/word2vec.

[33] J. Vlissides, R. Johnson, R. Helm, and E. Gamma. *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley Professional, 1994. URL https://www.safaribooksonline.com/library/view/design-patterns-elements/0201633612/?ar.

[34] Y. Xu, X. Guo, J. Hao, J. Ma, R. Y. Lau, and W. Xu. Combining social network and semantic concept analysis for personalized academic researcher recommendation. *Decision Support Systems*, 54, 2012. URL https://ac.els-cdn.com/S0167923612001996/1-s2.0-S0167923612001996-main.pdf?_tid=db93b81b-2679-4bef-a258-955706be71f5&acdnat=1540251311_015d33d3a2473a6c4a0cc9a9a2353146.

Appendix A

Data

Sample Dataset

An sample of data is shown below.

```
{"source": {
"title": "Towards ultra-stiff materials: Surface effects on
    nanoporous materials",
"author": [
"Dingjie",
"Yi Min",
"Qing",
"Xiaodong",
"Shiwei"
],
"abstract": "The significant rise in the strength and
   stiffness of porous materials at nanoscale cannot be
   described by conventional scaling laws. This letter
   investigates the effective Young's modulus of such
   materials by taking into account surface effect in a
   microcellular architecture designed for an ultralight
   material whose stiffness is an order of magnitude higher
    than most porous materials. We find that by considering
    the surface effects the predicted stiffness using Euler
   -Bernoulli beam theory compares well to experimental
   data for spongelike nanoporous gold with random
```

```
   microstructures. Analytical results show that, of the
   two factors influencing the effective Young's modulus,
   the residual stress is more important than the surface
   stiffness.",
"affiliations": [
"RMIT University",
"The University of Sydney"
],
"publisher": " American Institute of Physics Inc. subs@aip.
   org ",
"organization": "Applied Physics Letters",
"volume": "105"
}}
```

Appendix B

Hardware and Software

We use GitHub for version control.

- `https://github.sydney.edu.au/csun7801/Tinder4Researchers`

Hardware Requirements

Below are the hardware specifications used to host the web server.

- Model Name: MacBook Pro
- Processor Name: Intel Core i7
- Processor Speed: 2.8 Ghz
- Number of Processors: 1
- Total Number of Cores: 4
- Memory: 16 GB

Data Repository

The software uses Elasticsearch as data repository and Kibana as a plugin. Therefore, these components are required to install:

- `https://www.elastic.co/downloads/elasticsearch`
- `https://www.elastic.co/downloads/kibana`

A list of reference pages on how to configure Elasticsearch as well as Kibana can be found below:

- `https://www.elastic.co/guide/en/elasticsearch/reference/current/settings.html`
- `https://www.elastic.co/guide/en/elasticsearch/reference/2.3/setup-configuration.html`
- `https://www.elastic.co/guide/en/kibana/current/getting-started.html`

The communication between Elasticsearch and a back-end process is via a host address and a port number, which are `localhost` and `9200` respectively in a default setting. Changes of the host address and the port number of Elasticsearch will require to update `configure.py`.

```
HOST = "localhost"
PORT = "9200"
```

Back-End Process

The back-end process, which is responsible for distance computation, is written in Python. Therefore, its interpreter is required. The software requires minimum version of python as Python 3.6.4.

- `https://www.python.org/downloads/`

The required packages for Python are the following:

- `pip3 install elasticsearch`
- `pip3 install Counter`
- `pip3 install pandas`
- `pip3 install scipy`
- `pip3 install numpy`
- `pip3 install wget`
- `pip3 install scholarly`

- `pip3 install nltk`
- `pip3 install gensim`
- `pip3 install bz2file`
- `pip3 install newspaper3k`
- `pip3 install watchdog`
- `pip3 install pyemd`
- `pip3 install scipy`

Web Application

The front-end component, which is a data visualization interface, is written in a combination of HTML, Javascript and CSS. Because visuals are written in D3.js, which is in turn a Javascript library, a browser must support the HTML5 specification. However, we recommend Firefox.

- `https://www.mozilla.org/en-US/firefox/new/`

The server is written in Node.js, which is also a Javascript library.

- `https://nodejs.org/en/`

By installing Node.js, the npm package should have been installed. However, other required packages for Node.js are the following:

- `npm install elasticsearch`
- `npm install formidable`

Commands above must be run in the `_script` directory. Once they are executed, the `node_modules` directory and `package-lock.json` should be created.

Documents Ingest

Having installed the Elasticsearch, journals must be downloaded. `worker.py` helps ingesting journals from Scopus. To instruct what journals to download, an update to `Filtered-List-for-King.xlsx` is required in the `_meta` directory. Once it has been updated, the following command should be executed in the `_python` directory.

- `python3 worker.py`

Running Instructions

To execute the back-end process, the following command should be executed in the `_python` directory.

- `python3 author.py`

To start a server, the following command should be executed.

- `node _script/server.js 8080`

Finally, to view visuals is by typing this in the URL:

- `http://localhost:8080`

The evaluation results are stored in `_documents/results/evaluation_results.xlsx`. However, to re-evaluate models, the following command should be executed in the `_python` directory.

- `python3 evaluator.py`

Appendix C

Project Schedule

Prior to development, a detailed plan shown in the Table C.1 was drafted out to ensure a list of milestones along with scheduled and completion dates would be fulfilled. We mapped them out from the course outline. Dates were estimated based on experiences. We also identified the top priority was to complete the model implementation and evaluation. Nevertheless, it was understood visualisation had a dependency on the model implementation. The model evaluation surprisingly took longer than expected because we had troubles to find an appropriate baseline. We also had some throttle issues with Google Scholarly. Therefore, the network analysis part and Amazon Web Services were later de-scoped. The Bayesian statistics was definitely challenging to a person was previously considered himself a frequentist. The contingency option was to use the Frequentist inference. Fortunately, it did not happen. Challenges lied on technical capability too. For instance, a person who had little knowledges about web development had to learn Node.js, Angularjs and Elasticsearch within a short period of time. Certainly, on-line resources such as Coursera and Lynda helped. All in all, we were able to plan and follow the implementation process throughout the whole project despite the project complexities.

We adopted the Agile software development methodology in which requirements and solutions underwent a few iterative processes. Feedback from clients would be incorporated into a next release, and we have 6 releases thus far. This way, the software could be iteratively refined and improved until it met client expectations.

Milestones	**Descriptions**	**Status**	**Scheduled Date**	**Completion Date**
Project Kick-Off	Discuss with supervisors about the project's purposes, background, and requirements.	Completed	2^{nd}August 2017	2^{nd}August 2017
Literature Review	Research what has been written about this area. Identify strengths and weaknesses of each methodology.	Completed	30^{th}November 2017	30^{th}November 2017
Prototyping	Implement a prototype to demonstrate the methodology.	Completed	15^{th}January 2018	15^{th}January 2018
Presentation	Present a prototype to members from Faculty of Engineering and Information Technologies and Sydney Informatics Hub.	Completed	16^{th}January 2018	16^{th}January 2018
Project Proposal	Write a proposal about the project.	Completed	13^{th}April 2018	13^{th}April 2018
Implementation	Design and evaluate various distance metrics.	Completed	1^{st}June 2018	6^{th}September 2018
Progress Report 1	Write a progress report.	Completed	8^{th}June 2018	8^{th}June 2018
Product Development	Design and develop a web application until requirements are satisfied.	Completed	2^{nd}November 2018	2^{nd}November 2018
Progress Report 2	Write a progress report.	Completed	7^{th}September 2018	7^{th}September 2018
Literature review	Research on the bayesian statistics and generalised linear models.	Completed	19^{th}October 2018	19^{th}October 2018
Implementation	Design and evaluate various generalised linear models.	Completed	2^{nd}November 2018	18^{th}November 2018
Final Presentation	Present a final product to audience.	Completed	2^{nd}November 2018	2^{nd}November 2018
Final Report	Write a final report.	Completed	2^{nd}November 2018	18^{th}November 2018

Table C.1 Project Schedule